The Open University

M140

Introducing statistics

Handbook

The Open University

M140

Introducing statistics

Handbook

This publication forms part of the Open University module M140 *Introducing statistics*. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a catalogue (tel. +44 (0)1908 274066; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2014. Second edition 2015.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Page Bros, Norwich.

SUP 04373 8

2.1

Contents

Introduction

This Handbook is provided as a reference document for M140. It provides a concise summary of the material in the units, including the common notation and abbreviations used, a glossary of the important terms, and the key definitions and results.

Notation

Some of the notation used in M140 is listed below. Where notation has more than one possible meaning, the context should make it clear which meaning is intended in any given situation.

$\simeq$	approximately equal to
$>$	greater than
$\geq$	greater than or equal to
$<$	less than
$\leq$	less than or equal to
$\neq$	not equal to
$\sum$	sign indicating summation
$[+]$	the event that an observation is above the assumed population median
$[-]$	the event that an observation is below the assumed population median
$[=]$	the event that an observation is equal to the assumed population median
σ	population standard deviation
σ^2	population variance
μ	population mean
AP	the event that a child is initially allocated to the analytic phonics group
$AP+$	the event that a child is initially allocated to the analytic phonics + PA group
c	the number of categories for the variable corresponding to the columns in a contingency table
χ^2	the test statistic in a χ^2 test for contingency tables
${}^nC_x$	the number of ways of choosing x objects from a set of n objects if the order does not matter
CV1	critical value at the 1% significance level
CV5	critical value at the 5% significance level
df	degrees of freedom
E	expected value
ESE	an estimated standard error
E_L	lower extreme
E_U	upper extreme
f	frequency
HI	high outliers
H_0	null hypothesis of a hypothesis test
H_1	alternative hypothesis of a hypothesis test

IQR	interquartile range
LO	low outliers
M	population median
n	batch size
N	population size
O	observed value
p	price per unit of a commodity
$P(A)$	probability that the event A occurs
$P(A \text{ and } B)$	joint probability of A and B – that is, the probability that both A and B occur
$P(A\|B)$	conditional probability of A given B – that is, the probability that A occurs given that B occurs
P_{KS2}	average points score at the end of Key Stage 2
P_{KS4}	GCSE headline figure
q	quantity of a commodity
Q_1	lower quartile
Q_3	upper quartile
r	price ratio, *or* the number of categories for the variable corresponding to the rows in a contingency table, *or* the correlation coefficient
$R_0/R_1/R_2$	the event that reading age is higher than chronological age at the baseline/first follow-up test/second follow-up test
s	the (sample) standard deviation
s^2	the (sample) variance
s_{p}	the pooled estimate of the common population standard deviation
s_{p}^2	the pooled estimate of the common population variance
SE	a standard error – that is, the standard deviation of a sampling distribution
SP	the event that a child is initially allocated to the synthetic phonics group
$S_0/S_1/S_2$	the event that spelling age is higher than chronological age at the baseline/first follow-up test/second follow-up test
t	the test statistic in a t-test
t_{c}	the critical value in a t-test
w	weight (of some data)
$\overline{x}$	(arithmetic) mean
x_i	the ith data value in a batch of data
$x_{(i)}$	the ith data value in a batch of ordered data
$x[+]$	the event that x observations are above the assumed population median
$x[-]$	the event that x observations are below the assumed population median
z	an observation from the standard normal distribution, *or* the test statistic in a z-test

Abbreviations

Some of the abbreviations used in M140 are listed below.

AP	analytic phonics
APS	average points score
ASHE	Annual Survey of Hours and Earnings
AWE	Average Weekly Earnings
BAS	British Ability Scales
BCS	British Cohort Study
CHMP	Committee for Medicinal Products for Human Use
CPI	Consumer Prices Index
EMA	European Medicines Agency
GCSE	General Certificate of Secondary Education
iCMA	interactive computer-marked assignment
MWSS	Monthly Wages and Salaries Survey
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
OCAS	overall continuous assessment score
ONS	Office for National Statistics
PA	phonological awareness
PAYE	Pay-As-You-Earn
PRAC	Pharmacovigilance Risk Assessment Committee
RPI	Retail Prices Index
SATs	National Curriculum tests
SMC	Scottish Medicines Consortium
SP	synthetic phonics
TMA	tutor-marked assignment
VAT	value added tax
WRAT	Wide Range Achievement Test

Glossary

Below is a glossary of terms used in M140. The definition of each term is followed by a reference in brackets to the unit and section or subsection where the term is used. For example, '(U3, SS2.2)' refers to Subsection 2.2 of Unit 3.

adjacent value Adjacent values are the observations to which the whiskers on a boxplot are drawn. The lower adjacent value is the smallest data value within $Q_1 - 1.5 \times \text{IQR}$. The upper adjacent value is the largest data value within $Q_3 + 1.5 \times \text{IQR}$. (U3, SS2.2)

adverse event In a medical context, a treatment's side effect that has a negative effect on a patient. (U11, SS5.1)

all-commodities price ratio A weighted mean of price ratios for each commodity. The weights are the expenditures on each commodity in the previous year. (U2, SS4.1)

all-item price ratio A weighted mean of group price ratios. (U2, SS5.2)

alternative hypothesis The hypothesis that is assumed to be true if the null hypothesis is rejected. (U7, SS1.3)

analytic phonics A phonics approach to teaching reading that starts with learning whole words. (U8, SS1.1)

annual rate of inflation The percentage increase in CPI or RPI compared to one year earlier. Also known as the year-on-year rate of inflation, or just inflation. (U2, SS5.3)

arithmetic mean The arithmetic mean (or average) of a set of n numbers is the sum of the numbers divided by n. (U2, SS1.3)

Average Weekly Earnings (AWE) index An index which measures changes in earnings. (U3, SS5.1)

Baconian experiment See *exploratory experiment*.

base date The date at which an index is fixed to take a particular value (usually 100). (U2, SS4.1)

basket of goods The goods and services on which the RPI and CPI are based. (U2, SS5.1)

batch of data See *dataset*.

batch size The number of values in a batch of data. (U1, SS4.1)

bell curve See *normal distribution*.

bias In sampling, an error introduced by a poor sampling scheme. More generally, the effect on a measurement that tends to shift the results in a particular direction. (U4, SS4.1; U11, SS2.2)

bimodal See *mode*.

binomial distribution A distribution for the probability of the number of successes in a series of independent trials. (U12, SS3.2)

blind trial A clinical trial in which individuals do not know which treatment (one of which may be a placebo) they are given. A double-blind trial is one in which neither the doctors nor the patients know which treatment is being given. A single-blind trial is one in which the patients do not know which treatment they are given but the doctors do, or vice versa. (U11, SS2.2)

boxplot A diagram consisting of a box and whiskers, which displays the median, the quartiles and the minimum and maximum values of a batch of data. (U2, SS3.3; U3, SS2.2)

branch A line on a tree diagram showing a possibility (out of two or more) that exists at the start of the tree. (U6, SS2.3)

carry-over effect In crossover trials, any effect of a treatment that lasts beyond when it is given. (U11, SS3.1)

categorical data Data in which the observations are split into categories. Also known as nominal data. (U11, SS4.3)

categorical variable A variable for which the observed data can be classified into categories. (U8, SS2.2)

census A study in which data are collected from all the population. (U4, Introduction)

centre In clinical trials, a location (such as a clinic) where part of the trial is being held. (U11, SS3.4)

χ^2 **contribution** In the χ^2 test for contingency tables, the amount added to the test statistic relating to a cell in the contingency table. It is the quantity $(\text{Observed} - \text{Expected})^2/\text{Expected}$. (U8, SS4.3)

χ^2 **family of distributions** A particular family of models for the variation in a variable. It assumes that the variable cannot take negative values. (U8, SS4.4)

χ^2 **test for contingency tables** A hypothesis test in which the hypothesis that two categorical variables are statistically independent is tested. (U8, S4)

chronic condition A medical condition (such as an illness) that is relatively long-lasting. (U11, SS3.1)

cleaning the data The process of preparing data ready for analysis. (U1, SS3.1)

clinical trial A study in which a new treatment is compared with an existing therapy or, sometimes, compared with what happens when no treatment is given. (U11, SS1.1)

cluster The population in a distinct geographical area or other similar subset of the population. (U4, SS4.3)

cluster sampling A method of sampling where first some clusters are selected, and then only individuals from the selected clusters are sampled. (U4, SS4.3)

cohort In clinical trials, a group of patients followed up over time. (U11, SS5.3)

combination A selection from a set of items where the order in which the items are selected does not matter. (U6, SS3.1)

commodity An item or service that people pay for. (U2, SS2.2)

common population variance When the variances of two populations are assumed to be equal, the common population variance is the value these two variances take. This quantity is usually estimated from data to give a 'pooled estimate' of the common population variance. (U10, SS3.3)

complementary events Two events are complementary when one, and only one, of the events can occur at any given time. (U6, SS2.2)

conditional probability The conditional probability of A given B is the probability that A occurs when it is known that B has occurred. (U8, SS3.1)

confidence interval An interval estimate for a population value, consisting of all hypothesised values that cannot be rejected at a given significance level. (U9, SS4.2; U9, SS5.1)

confounding The relationship between two variables, A and B, is said to be confounded by a third variable, C, if the apparent association between A and B is distorted due to the associations of A and B with C. (U8, SS5.1)

Consumer Prices Index (CPI) An index used to record changes in the average level of prices that most people pay for good and services. (U2, SS5.1)

contingency table A table of counts where the rows correspond to categories of one variable and the columns correspond to categories of a different variable. When the variable used for the rows has r categories and the variable used for the columns has c categories, the table is also known as an $r \times c$ contingency table. (U8, SS2.2)

control group A group of subjects in a clinical trial, or other experiment, who do not receive the treatment being tested. (U11, SS2.1)

coordinates A pair of numbers indicating a position on a scatterplot. (U5, SS1.3)

correlation coefficient A number which summarises the strength of a (linear) relationship between two variables. (U9, SS2.1)

critical region (at a given significance level) All the values of the test statistic that would lead to the hypothesis being rejected. It corresponds to the 'most extreme' values of the test statistic. (U6, SS4.1; U7, SS5.1)

critical value (at a given significance level) The least extreme value of the test statistic for which the hypothesis would be rejected. In other words it is an 'inner end' of the critical region. (U6, SS4.1; U7, SS5.1)

crossover trial A clinical trial where patients receive the experimental treatment followed by the control treatment, or vice versa. (U11, SS3.1)

cumulative probability A probability of the form $P(X \leq A)$ for some value A. (Computer Book, SS6.1)

data Observations, usually numbers. (U1, SS2.1)

data collection The process of making observations. (U1, SS2.1)

dataset A collection of observations – also known as a 'batch of data'. (U1, SS2.1)

decile A percentile for which the corresponding percentage is a multiple of 10. The 10th percentile is the lowest decile. The 90th percentile is the highest decile. (U3, SS1.5)

decile boxplot A boxplot which extends only from the lowest decile to the highest decile. The ends of the whiskers are usually denoted by arrowheads. (U3, SS1.5)

degrees of freedom A value which indicates the exact distribution in the family of χ^2 distributions or the family of t distributions. (U8, SS4.4; U10, SS3.3)

dependent variable See *response variable*.

design (of an experiment) Factors such as the nature of the control group and the kinds of measurement taken, that collectively make up an experiment. (U11, SS2.2)

deviation (of a data value from the mean) The deviation of a value x from the sample mean $\overline{x}$ is $x - \overline{x}$. (U3, SS3.1)

DFR equation An equation that gives the link between data, fitted values and residuals. (U5, SS3.2)

dimension (of a contingency table) The number of rows and the number of columns in a contingency table. Also known as the size of the contingency table. (U8, SS2.2)

distribution The shape and location of a dataset. (U1, SS4.1)

double-blind trial See *blind trial*.

earnings distribution The distribution of earnings across a group of employees. (U3, SS1.4)

earnings ratio The ratio of women's earnings to men's earnings. Usually calculated at a particular point in the distribution such as the mean, median, quartiles or deciles. (U3, SS1.2, SS1.4, SS1.6)

efficient A sampling method is said to be efficient if it has a relatively small sampling error. (U4, SS4.2)

empirical data Data collected through observation or experimentation. (U10, Introduction)

error The difference between a population value (such as the median) and the value obtained from a sample. (U4, SS4.1)

exact relationship A relationship where, on a scatterplot, all the points lie exactly on a straight line or follow a simple curve. (U5, SS3.1)

Expected value When carrying out the χ^2 test for contingency tables, the Expected values correspond to the counts that would be expected to occur if the two variables are independent. (U8, SS4.2)

experiment An activity that stands up to scrutiny, is repeatable and sets out to answer a specific question or set of questions. (U10, SS1.1)

experimental group The group of subjects in a clinical trial, or other experiment, who receive the experimental treatment. (U11, SS2.1)

explanatory variable When investigating the relationship between two variables, this is the variable that is doing the explaining or the variable on which the response variable depends. Also known as the independent variable. (U5, SS1.4)

exploratory experiment An experiment that is exploratory in nature. Also known as a Baconian experiment. (U10, SS1.2)

extrapolation Making predictions outside the range of values in the original sample. (U12, SS6.2)

first coordinate See *x-coordinate*.

first quartile See *quartile*.

fit value The value of the response variable which, according to a model, corresponds to a given value of the explanatory variable. (U5, SS3.2)

fitting a line The process of choosing a straight line to represent a relationship. (U5, SS3.1)

fitting by eye Fitting a line to data on the basis that it looks to be a good representation of the relationship. (U5, SS3.1)

five-figure summary A summary of a batch of data giving the minimum, lower quartile, median, upper quartile and maximum. (U2, SS3.3)

frequency (of a particular value) The number of times that the value occurs. (U3, SS3.2)

Gaussian distribution See *normal distribution*.

GCSE headline figure A school's GCSE headline figure is the percentage of students ending Key Stage 4 who achieve at least five grade A* to C GCSEs, including English and Mathematics. (U9, SS1.1)

gender differential (in pay) The difference in the amount that men and women earn. (U3, Introduction)

grapheme A letter, or combination of letters, of the alphabet, corresponding to an individual sound. (U8, SS1.1)

gross earnings Earnings before deductions (such as tax, pension and national insurance). (U3, SS1.2)

group-comparative trial A clinical trial in which patients are split into groups, with the intention that each group is representative of the patient population. Treatments (experimental or control) are then allocated to the groups so that each member of a group receives the same treatment. (U11, SS3.3)

grouped data Data which is split into groups with the number of observations in each group recorded. (U3, SS3.2)

growth chart A chart that displays the distribution of weights or heights across a range of ages. (U12, SS1.2)

high outlier See *outlier*.

highest decile See *decile*.

histogram A diagram that represents a dataset in which a range of values is represented by a rectangle whose length is proportional to the frequency of that range of values. (Computer Book, SS1.5)

hypothesis test A procedure where sample data are used to evaluate the credibility of a statement. (U6, S4)

hypothesis-testing experiment An experiment designed to see if predictions that flow from a hypothesis are correct. Also known as a hypothetico-deductive experiment. (U10, SS1.2)

hypothetico-deductive experiment See *hypothesis-testing experiment*.

hypothetico-deductive method Using scientific experimentation to test deductions that can be made from a hypothesis. (U10, SS1.3)

independence See *statistical independence*.

independent variable See *explanatory variable*.

index-linking (or indexation) The adjustment of an amount of money (such as a pension) by the same ratio as the change in the value of a price index (such as the RPI or CPI). (U2, SS5.3)

inference See *statistical inference*.

inflation See *annual rate of inflation*.

influential point A point on a scatterplot which follows the pattern of the other points, but is a long way from most of them; hence it has a strong influence on the correlation. (U9, SS3.3)

informed consent The agreement given by a person being treated once they know all the relevant information. (U11, SS2.1)

interquartile range (IQR) The distance between the upper quartile and the lower quartile of a batch of data. (U2, SS3.2)

interval A range of numbers between which a quantity is likely to lie. (U9, SS4.1)

interval estimate An estimate of a population value (such as the population mean) from a sample, specified by a range (interval) of likely values. (U9, SS4.1)

interval scale data Data that correspond to measurements where the difference between two values is meaningful. (U11, SS4.3)

joint probability The probability that two (or more) events occur together. (U8, SS3.1)

leaf (of a stemplot) A single digit on the right-hand side of a stemplot that represents a single data value. (U1, SS4.1)

least squares A method for fitting a line to a set of data points in such a way that the sum of the squared residuals is as small as possible. (U5, SS4.1)

least squares regression line (or least squares fit line) The straight line fitted using the method of least squares. It is often referred to as simply the 'regression line'. (U5, SS4.1, SS4.2)

left-skew See *skew*.

level (of a stemplot) A row of a stemplot. (U1, SS4.1)

Likert scale A scale consisting of a number of categories, one of which a respondent selects to indicate their level of agreement or disagreement with a statement. (U4, SS3.1)

linear regression See *regression*.

linear relationship A relationship between two variables that can be summarised reasonably well by a straight line. (U5, SS2.2)

linked data Two or more variables that are recorded for the same sampling units. Also known as paired data when there are just two variables. (U5, SS1.2)

location (measure of) A quantity that gives a typical value of a dataset. Measures of location include the (arithmetic) mean and median. (U1, SS4.2; U2, SS1.4)

low outlier See *outlier*.

lower adjacent value See *adjacent value*.

lower extreme The smallest value in a dataset. (U1, SS4.2)

lower quartile See *quartile*.

lowest decile See *decile*.

marginal totals The row and column totals of a table. (U8, SS2.1)

matched pairs Pairs of individuals who have been matched by particular features or diseases likely to be important to the outcome of any treatment. (U10, SS4.2; U11, SS3.2)

matched-pairs trial A clinical trial in which the participating patients are split into matched pairs. Then, in each pair, one patient is given the experimental treatment and the other is given the control treatment. (U11, SS3.2)

matched-pairs t-test A hypothesis test, based on the t distribution, used to test the hypothesis that the difference between two population means is zero from matched-pairs data. It is equivalent to a one-sample t-test based on the differences when the hypothesised mean is equal to zero. (U10, SS4.2)

mean (of a dataset) See *arithmetic mean*.

measurement experiment An experiment with the primary purpose being the measurement of a particular attribute. (U10, SS1.2)

median If the values in a batch of data are written in order of increasing size, then the median is the middle value when the batch size is odd, or the average of the middle two values when the batch size is even. The median is also known as the 'middle' of the batch. (U1, SS4.2)

median response The median of all responses from a population or sample. (U4, SS3.2, SS3.3)

middle (of a batch) See *median*.

mid-range The average of the minimum and maximum in a batch of data. It is a measure of location. (U12, SS1.2)

mode A clear peak in a distribution. A distribution that has only one clear peak is called unimodal; a distribution that has two clear peaks is called bimodal; a distribution that has more than two clear peaks is called multimodal. (U1, SS5.2)

model A mathematical description used to describe the relationships underlying a dataset. (U1, SS2.1)

modelling diagram A diagram showing the modelling process – that is, the process of posing a question, collecting data, analysing data and interpreting the results. (U1, SS2.1)

multimodal See *mode*.

mutually exclusive Two or more events are said to be mutually exclusive if no two of them can occur at the same time. (U6, SS2.2)

negative relationship A relationship between two variables where low values of one variable are associated with high values of the other variable. (U5, SS2.1)

nominal data See *categorical data*.

non-linear relationship A relationship that can be summarised reasonably well by a curve but not by a straight line. (U5, SS2.2)

non-sampling error The error associated with a sample result that is not simply due to the fact that different samples contain different individuals. (U4, SS4.1)

normal distribution A particular model for the variation in a variable. Also known as the Gaussian distribution, and the 'bell-curve'. (U7, S2; U7, SS3.1)

null hypothesis The hypothesis about a population (or populations) that is initially assumed to be true, but which may or may not be rejected as the result of a hypothesis test. (U7, SS1.3)

objective measurement A measurement such as blood pressure or pulse rate, that is measured in an objective way. (U11, SS2.2)

Observed value When carrying out the χ^2 test for contingency tables, the Observed values correspond to the counts given in the table. (U8, SS4.2)

one-sample *t*-test See *t-test*.

one-sample *z*-test See *z-test*.

one-sided alternative hypothesis An alternative hypothesis that states that the population quantity is bigger than a particular value, for example $\mu > A$, or an alternative hypothesis that states that the population quantity is smaller than a particular value, for example $\mu < A$. (U10, S6)

one-sided (hypothesis) test A hypothesis test that uses a one-sided alternative hypothesis. (U10, S6)

ordinal data Data that are on an ordered scale and only the ordering is meaningful. (U11, SS4.3)

outlier An observation that is noticeably separated from, or inconsistent with, the rest of the data. If it is greater than the rest of the data, then it is known as a high outlier; if it is less than the rest of the data, then it is known as a low outlier. (U1, SS4.2; U5, SS2.4)

***p*-value** The *p*-value (or significance probability) is the probability that a value of the test statistic occurs which is at least as extreme as the one observed if the null hypothesis is true. (U6, SS5.1)

paired data See *linked data*.

patient-years A measurement of exposure to a particular treatment, such as a drug. One patient-year is equivalent to one person receiving the treatment for one year. (U11, SS5.2)

percentile The nth percentile is the number such that $n\%$ of the values in the batch fall below it. (U3, SS1.5)

phoneme An individual sound, relating to a letter or group of letters, that can distinguish one word from another. (U8, SS1.1)

phonics Grapheme–phoneme, or letter–sound, correspondences. (U8, SS1.1)

placebo A dummy treatment that superficially resembles the treatment being tested, but contains no active ingredient. (U11, SS2.1)

placebo-controlled trial A clinical trial in which the control group are given a placebo. (U11, SS2.1)

placebo effect The apparent therapeutic effect from a treatment that contains no medication and hence should be ineffectual. (U11, SS2.1)

point estimate An estimate of a population value (such as the population mean) from a sample, which consists of a single number. (U9, SS4.1)

pooled estimate of the common population variance See *common population variance.*

population The collection of all the individual values or members of a specified group of interest. (U4, Introduction)

population distribution The shape and location of values in all the population. (U4, SS3.4)

population least squares regression line The regression line that would be calculated if all values in the population were known for both variables. The slope of this line is known as the 'population slope' and the intercept is known as the 'population intercept'. (U9, SS5.1)

population mean The mean of all the values in a population. (U7, S3)

population median The median of all the values in a population. (U4, SS3.2)

population standard deviation The standard deviation of all the values in a population. (U7, S3)

population values The values of a variable in the whole population. For example, the responses all members in a population would give to a question. (U4, SS3.1)

positive relationship A relationship between two variables where low values of one variable are associated with low values of the other variable, while high values of one variable are associated with high values of the other variable. (U5, SS2.1)

powerful (test) A test is said to be more powerful than another if it is better at identifying a null hypothesis that is false. (U10, SS4.1)

predicted value The forecasted value of an observation or measurement, based on a model fitted to the data. (U5, SS5.2)

prediction The process of forecasting a value of the response variable for a given value of the explanatory variable. (U1, SS2.1; U5, SS5.2)

prediction interval An interval estimate that reflects the random variation of individual values around the population regression line, as well as uncertainty about the actual position of that line. (U9, SS5.2)

price ratio The price of a commodity in one year divided by its price in the previous year. (U2, SS4.1)

probability The probability of an event is the likelihood or chance that the event occurs. (U6, S2)

probability distribution The set of probabilities relating to all possible outcomes for a variable. That is, the probabilities of all possible events that could occur. (U6, SS3.2)

purchasing power (of the pound) The amount a consumer can buy with a fixed amount of money at one point of time compared with another point of time. (U2, SS5.3)

quartile The median and quartiles of a batch of data divide the batch into four roughly equal parts. Roughly 25% of the values in the batch are smaller than the lower (first) quartile (Q_1), and roughly 25% of the values are greater than the upper (third) quartile (Q_3). (U2, SS3.2)

quota sampling A sampling method in which interviewers are sent out to selected sites to interview a fixed number of people in specified groups. (U4, SS4.5)

$r \times c$ contingency table See *contingency table.*

randomisation In clinical trials, the process of using methods based on random numbers to allocate patients to treatments. (U11, SS3.4)

random sampling (or selection) A sampling method in which every possible sample has an equal chance of being selected. (U4, SS1.2)

random start In systematic sampling, the first label selected (at random). (U4, SS2.2)

range The range of a batch of data is the difference between the maximum and minimum values. (U1, SS4.2)

real earnings (for a particular month compared with one year earlier) The relative amount a person's earnings can buy compared with their earnings one year earlier (after allowing for inflation). (U3, SS5.2)

record linkage Using records such as a patient's medical record to look for links, other than those already known, between various ailments and the drugs being taken. (U11, SS5.3)

regression The process of finding a line that best represents a relationship. When the line is a straight line it is also known as linear regression. (U5, SS3.1, SS4.2)

regression line See *least squares regression line.*

related variables Two variables are related if knowing the value of one variable provides information about the value of the other variable. In that case, there is said to be a relationship between them. (U5, S1)

relationship See *related variables.*

remote point A point on a scatterplot that is a long way from most other points. (U9, SS3.3)

repeatable (experiment) Suppose person A carries out an experiment, then the experiment is repeatable if they are able to explain everything that took place in such a way that another person (B) could, if necessary, go through exactly the same procedure. (U10, SS1.1)

residual When a line is fitted to a set of data points, the residual of each data pair may be calculated using the relationship Residual = Data − Fit, where Data is the y-coordinate of the data pair and Fit is the y-value predicted by the line for the corresponding x-coordinate. (U5, SS3.2)

residual plot A scatterplot where the horizontal axis represents the explanatory variable and the vertical axis represents the value of the residuals. (U5, SS3.3)

resistant measure A measure which is insensitive to changes in the values near the extremes. (U2, SS1.4)

response variable A variable that is being explained or whose value depends on other variables. It is also the variable to be predicted if predictions are to be made. Also known as the dependent variable. (U5, SS1.4)

Retail Prices Index (RPI) An index used to record changes in the average level of prices that most people pay for good and services. (U2, SS5.1)

right-skew See *skew*.

rounding When a value is given to a specified degree of accuracy it is said to be rounded. (U1, SS3.2)

rounding error The difference between a value calculated using full accuracy and a value calculated using rounded values. (U1, SS3.2)

sample A subset of a population. (U4, Introduction)

sample data The data obtained from a sample. (U4, SS3.1)

sample median The median of a sample. (U4, SS3.3)

sampling distribution of the difference between two means The distribution of the difference between means of random samples from two populations. (U7, S6)

sampling distribution of the mean The distribution of the mean of random samples of size n from a population. (U7, S2)

sampling distribution of the median The distribution of the medians of all possible samples of size n from a population is called the sampling distribution of the median for samples of size n. (U4, SS3.3)

sampling error Variability that is due to sampling. (U4, SS4.1)

sampling frame A list of all the members in a target population. (U4, SS4.6)

sampling interval In systematic random sampling, the gap between successive sampled individuals in the sampling frame. (U4, SS2.2)

scatter See *spread*.

scatterplot A graph used to represent linked data. On a scatterplot each observation is represented by a point, its position on the x-axis indicating

its value for the explanatory variable and its position on the y-axis indicating its value for the response variable. (U1, SS2.1; U5, SS1.3)

scientific experiment See *experiment*.

second coordinate See *y-coordinate*.

selection bias In a clinical trial, bias that arises due to the way that patients were selected to receive either the experimental treatment or the control treatment. (U11, SS3.4)

sensitive measure A measure is said to be sensitive if its value can be substantially influenced by outliers. (U2, SS1.4)

side effects Any effects (usually unwanted) of drugs or other treatments, additional to that intended. (U11, SS5.1)

sign test A hypothesis test used to test the hypothesis that a population median has a particular value. (U6, SS4.1, SS4.2)

significance level (of a hypothesis test) The point at which the null hypothesis should be rejected as the probability (or percentage) of the sample occurring when the null hypothesis is true is deemed sufficiently small. (U6, SS4.1)

significance probability See *p-value*.

significant figures The first significant figure of an unrounded number is its first non-zero digit, counting from the left. The next significant figure is the next digit (zero or other), and so on. (U1, SS3.3)

simple random sampling A sampling method in which every sample of size n has an equal chance of being selected. (U4, SS1.3)

single-blind trial See *blind trial*.

size (of a contingency table) See *dimension (of a contingency table)*.

skew A dataset which is not symmetric is said to be skew. If the large data values are more spread out than the small data values, then the dataset is right-skew. If the large data values are less spread out than the small data values, then the dataset is left-skew. (U1, SS5.2; U3, SS2.1)

sources of variation In a clinical trial, factors that are known to lead to differences in measurements between patients. (U11, SS2.3)

spread (measure of) A quantity that indicates how spread out or scattered values in a dataset are. Measures of spread include the range, interquartile range, variance and standard deviation. (U1, SS4.2; U2, S3)

spurious accuracy The display of data to more decimal places or significant figures than is justified from the context. Also known as spurious precision. (U1, SS3.1)

spurious precision See *spurious accuracy*.

standard deviation A measure of spread that is a kind of average deviation. It is the square root of the variance. (U3, SS3.1)

standard error of the difference between two means The standard deviation of the sampling distribution of the difference between two means. (U7, S6)

standard error (of the mean) The standard deviation of the sampling distribution of the mean. It is the population standard deviation divided by the square root of the sample size. (U7, S4)

standard normal distribution The normal distribution that has mean $\mu = 0$ and standard deviation $\sigma = 1$. (U7, SS3.3)

statistical independence Two events are said to be statistically independent if the occurrence of one event has no effect on the likelihood of occurrence of the other event. (U6, SS2.3)

statistical inference The process of inferring back from a sample to the population. (U6, S1, SS1.2)

stem (of a stemplot) A column of numbers on the left hand side of the stemplot indicating the levels used. (U1, SS4.1)

stemplot A plot for displaying numerical data. It consists of three elements: the stem, levels and leaves. (U1, SS4.1)

strata Categories into which a population is divided. (A single category is called a stratum.) In clinical trials, strata often correspond to age, centres, time periods or disease progression. (U4, SS4.2; U11, SS3.4)

stratification The process of dividing a target population into distinct categories. (U4, SS4.2)

stratified randomisation This occurs in clinical trials when the randomisation for each stratum is carried out separately. (U11, SS3.4)

stratified sample A sample that consists of subsamples taken from each stratum in the population. (U4, SS4.2)

stratum See *strata*.

stretched stemplot A stretched stemplot is one in which levels are split into more than one part, each part allowing different possible digits for the leaves. (U1, SS5.1)

strong relationship A relationship where all the points on a scatterplot lie close to a line. (U5, SS2.3)

sub-batch Part of a batch of data. (U2, SS1.2)

sub-branch A line on a tree diagram showing a possibility that exists, following on from a previous branch. (U6, SS2.3)

subjective measurement A measurement such as degree of pain, that is measured in a subjective way. (U11, SS2.2)

subsample Part of a sample. (U4, SS4.2)

summary measures Quantities that summarise aspects of a batch, such as its location and spread. (U3, S3)

survey A study in which information is gathered from a sample of the target population. (U4, Introduction)

symmetric A dataset is symmetric when the shape of the values below the median mirrors the shape of the values above the median. (U1, SS5.2)

synthetic phonics An approach to teaching reading that starts by teaching letter sounds, and then building them up into whole words. (U8, SS1.1)

systematic error An error in an experiment that is consistent in its direction and approximately constant in its magnitude. For example, the underestimation resulting from trying to measure curved plant roots against a straight ruler. (U10, SS2.7)

systematic random sampling A sampling method where only the first individual in the sample is selected at random. The remaining individuals in the sample are selected due to their position at successive fixed intervals in the sampling frame. (U4, SS2.2)

t distribution A particular model for the variation in a variable. It is similar in shape to the standard normal distribution, however the probability of getting an extreme value (positive or negative) increases as the number of degrees of freedom decreases. (U10, SS3.3)

t-test A hypothesis test, based on the t distribution, used to test the hypothesis that a population mean takes a particular value (the one-sample version) or that two population means are equal (the two-sample version). (U10, SS3.3, SS4.1)

tails of a batch The values away from the largest mode. (U1, SS5.2)

target population The population of interest in a survey. (U4, S1)

test statistic A quantity that is calculated from data and can be used to decide whether or not to reject the null hypothesis. For example, in the sign test it is the smaller of the number of [+]s and [−]s. (U7, SS1.3)

third quartile See *quartile*.

tie In the sign test, a tie is the situation when a sample value is equal to the assumed population median. (U6, SS5.2)

tree A diagram formed by branches and sub-branches that together represent combinations of events. (U6, SS2.3)

truancy rate In M140, this is taken to be the unauthorised absence rate from school. A pupil's truancy rate is the proportion of school half-days that the pupil was absent without authorisation. A school's truancy rate is the average truancy rate of its pupils. (U6, SS1.2)

truncated number A number is truncated when its last digits have been dropped without rounding. (U1, SS4.2)

two-sample z-test See *z-test*.

two-sided alternative hypothesis An alternative hypothesis in which a population quantity is assumed not to take a particular value. For example, $\mu \neq A$. (U7, S5)

two-sided (hypothesis) test A hypothesis test that uses a two-sided alternative hypothesis. (U10, S6)

type 1 error This type of error occurs when the null hypothesis is rejected even though it is true. (U8, SS5.2)

type 2 error This type of error occurs when the null hypothesis is not rejected even though it is false. (U8, SS5.2)

unauthorised absence rate See *truancy rate*.

unimodal See *mode*.

unordered stemplot A stemplot where the leaves on each level are not given in a particular order. (U1, SS4.1)

upper adjacent value See *adjacent value*.

upper extreme The largest value in a dataset. (U1, SS4.2)

upper quartile See *quartile*.

variable A quantity that varies from one sampling unit to another. (U5, S1)

variance A measure of spread. It is the square of the standard deviation. (U3, SS3.1)

weak relationship A relationship where the points on a scatterplot only loosely follow a line. (U5, SS2.3)

weight A quantity indicating the degree of importance to be given to a data value in a calculation. (U2, SS2.1)

weighted mean A mean where data values are not all given the same weight in the calculation. (U2, SS2.1)

x-axis The horizontal axis on a plot. (U5, SS1.3)

x-coordinate The first value given in a pair of coordinates. It corresponds to the position along the x-axis. Also known as the first coordinate. (U5, SS1.3)

y-axis The vertical axis on a plot. (U5, SS1.3)

y-coordinate The second value given in a pair of coordinates. It corresponds to the position along the y-axis. Also known as the second coordinate. (U5, SS1.3)

year-on-year rate of inflation See *annual rate of inflation*.

z-test A hypothesis test, based on the normal distribution, which is used to test the null hypothesis that a population mean takes a particular value (the one-sample version) or that two population means are equal (the two-sample version). (U7, S5, S6)

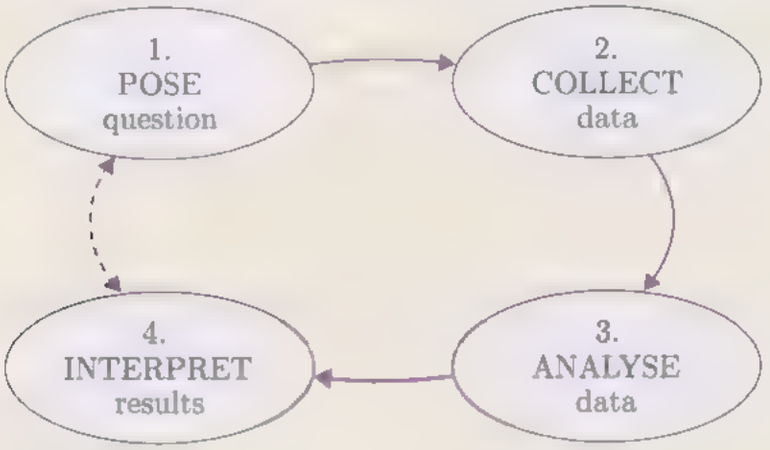
Definitions and results

The following definitions and results have been collected from the units of M140.

Unit 1 Looking for patterns

Modelling process

The process of modelling data can be split into four stages. These stages are summarised in the following diagram.



The modelling diagram

Rounding

- Numbers should always be quoted to a suitable degree of accuracy to avoid giving a false impression ('spurious accuracy' or 'spurious precision').
- To round a number, find the digit immediately to the right of where you want to round.

  Round up if this digit is 5 or more, and round down otherwise.
- To reduce rounding errors in the final result, the full accuracy available should be kept in intermediate calculations, and the result should be rounded at the end.
- The 'output' result of a multiplication or division of several 'input' quantities, some of which have been rounded, should be rounded so that it has the same number of significant figures as the rounded input quantity with the smallest number of significant figures.

Stemplots

- The stemplot is a device for displaying numerical data in a pictorial structure. The three basic elements in the stemplot are the stem, the levels and the leaves.
- The stem of a stemplot is the single column of figures lying to the left of the vertical line, arranged vertically downwards in increasing order. This means that the stemplot breaks up the range of data values into a column (stem) of ordered levels corresponding to convenient equal intervals.
- The level of a stemplot consists of everything occurring on the same horizontal line of the stemplot as the number used to name it. The levels of a stemplot are labelled so that we know what each interval is.
- Each leaf of a stemplot represents a data value whose first few significant figures are the level followed by the value for the leaf. For example, leaf b at level a represents a data value for which the significant figures start 'ab'. The leaves on a level are usually given in numerical order.
- The number of leaves at each level tells us the number of data values in each interval. This indicates if there are any 'gaps' without many values, and where the maximum and minimum values are. Comparing the numbers of leaves at the various levels indicates whether the data values 'cluster together' in particular regions.
- The number of levels in a stemplot should be chosen so that it is neither too cramped nor too spread out. The number of levels in a stemplot can be increased by stretching it – splitting each level into parts. The number of levels in a stemplot can be reduced by squeezing it – combining levels. The number of levels in a stemplot can also be reduced by listing outliers separately.

Shape of data

The stemplot can be used to picture the shape of a batch of data, in particular the number of modes (peaks) and whether there is any symmetry.

- One clear peak in the stemplot indicates that a batch of data is unimodal.
- Two clear peaks in the stemplot indicate that a batch of data is bimodal.
- Three or more clear peaks indicate that a batch of data is multimodal.
- A batch of data is symmetric if a horizontal line can be drawn across a stemplot so that the shape on one side is a mirror image of the other.
- A batch of data is right-skew if large values are more spread out than small values.

- A batch of data is left-skew if small values are more spread out than large values.

Right-skew Left-skew

Stemplot shapes

Median

The median of a batch of data is the middle of the ordered batch. It is a resistant measure. That is, its value is not much influenced by the presence of outliers.

- If the batch size is odd.

  median = middle data value.

- If the batch size is even.

  $$\text{median} = \text{average of the two middle data values} = \frac{\text{sum of the two middle data values}}{2}.$$

The median should be rounded to the same level of accuracy as the original data.

Range

The range of a batch is the distance between the two extreme values. It can be calculated from the formula

$$\text{range} = E_U - E_L,$$

where E_U is the upper extreme (the largest value, or maximum) and E_L is the lower extreme (the smallest value, or minimum).

Unit 2 Prices

Properties of sub-batches

Two general properties of sub-batches are as follows.

- The range of the complete batch is greater than or equal to the ranges of all the sub-batches.
- The median of the complete batch is greater than or equal to the smallest median of a sub-batch and less than or equal to the largest median of a sub-batch.

Arithmetic mean

The arithmetic mean is the sum of all the values in the batch divided by the size of the batch. More briefly,

$$\text{mean} = \bar{x} = \frac{\text{sum}}{\text{size}} = \frac{\sum x}{n}.$$

The mean is a sensitive measure.

Weighted means

- The formula for the mean $\bar{x}_C$ of a combined batch C is

  $$\bar{x}_C = \frac{\bar{x}_A n_A + \bar{x}_B n_B}{n_A + n_B}.$$

  where batch C consists of batch A combined with batch B, and $\bar{x}_A$ = mean of batch A, n_A = size of batch A, $\bar{x}_B$ = mean of batch B, n_B = size of batch B.
- In general, if you purchase q_1 units of some commodity at p_1 pence per unit, and q_2 units of the same commodity at p_2 pence per unit, then the mean price of this commodity, $\bar{p}$ pence per unit, can be calculated from the following formula:

  $$\bar{p} = \frac{p_1 q_1 + p_2 q_2}{q_1 + q_2}.$$
- The weighted mean of the two numbers x_1 and x_2 with corresponding weights w_1 and w_2 is

  $$\frac{x_1 w_1 + x_2 w_2}{w_1 + w_2}.$$
- The weighted mean of two or more numbers is

  $$\frac{\text{sum of } \{\text{number} \times \text{weight}\}}{\text{sum of weights}} = \frac{\text{sum of products}}{\text{sum of weights}}.$$
- The following three rules apply to weighted means.

Rule 1 The weighted mean depends on the relative sizes (i.e. the ratio) of the weights.

Rule 2 The weighted mean of two numbers always lies between the numbers and it is nearer the number that has the larger weight.

Rule 3 If the weights are equal, then the weighted mean of two numbers is the number halfway between them.

The quartiles

The lower quartile, Q_1, is at position $\frac{(n+1)}{4}$ in the ordered batch.

The upper quartile, Q_3, is at position $\frac{3(n+1)}{4}$ in the ordered batch.

If $(n+1)$ is exactly divisible by 4, these positions correspond to a single value in the batch.

If $(n+1)$ is not exactly divisible by 4, then the positions are to be interpreted as follows.

- A position which is a whole number followed by $\frac{1}{2}$ means 'halfway between the two positions either side' (as was the case for finding the median).
- A position which is a whole number followed by $\frac{1}{4}$ means 'one quarter of the way from the position below to the position above'. So for instance if a position is $5\frac{1}{4}$, the quartile is the number one quarter of the way from $x_{(5)}$ to $x_{(6)}$.
- A position which is a whole number followed by $\frac{3}{4}$ means 'three quarters of the way from the position below to the position above'. So for instance if a position is $4\frac{3}{4}$, the quartile is the number three quarters of the way from $x_{(4)}$ to $x_{(5)}$.

The interquartile range

The interquartile range (sometimes abbreviated to IQR) is the distance between the lower and upper quartiles:

$$\text{IQR} = Q_3 - Q_1.$$

Five-figure summary

For a batch of data, the five-figure summary is as follows.

		M	
n	Q_1		Q_3
	E_L		E_U

n batch size
M median
Q_1 lower quartile
Q_3 upper quartile
E_L lower extreme
E_U upper extreme

The five-figure summary can be displayed pictorially using a boxplot. (For more details about boxplots, see the Unit 3 summary.)

Procedure: calculating a chained price index

1. For each year calculate the following.
   - The price ratio for each commodity covered by the index:
     $$\frac{\text{price that year}}{\text{price previous year}}.$$
   - The weighted mean of all these price ratios, using as weights the expenditure on each commodity in the previous year. This weighted mean is called the all-commodities price ratio.
2. For each year, the value of the index is

   value of index for previous year × all-commodities price ratio.

   The value of the index in the first year is set at 100; this date is the base date of the index.

The Retail Prices Index (RPI)

The Retail Prices Index (RPI), along with the Consumer Prices Index (CPI), is a measure used in the UK to record changes in the average level of the prices most people pay for the goods and services they buy. The RPI is intended to reflect the average spending pattern of the great majority of private households.

The bulk of the data on price changes required to calculate the RPI is collected by staff of a market research company and forwarded to the Office for National Statistics for processing.

The calculation for the RPI is as follows.

1. The data used are prices, collected monthly, and weights, based on the Living Costs and Food Survey, changed annually.
2. Each month, for each item, a price ratio is calculated, which gives the price of the item that month divided by its price the previous January.
3. The group price ratios are calculated from the price ratios using weighted means.
4. Weighted means are then used to calculate the all-item price ratio. Denoting the group price ratios by r and the group weights by w, the all-item price ratio is
   $$\frac{\sum rw}{\sum w}$$
5. The value of the RPI for that month is found by multiplying the value of the RPI for the previous January by the all-item price ratio:

   RPI for month x = RPI for previous January
   × (all-item price ratio for month x).

The weights for a particular year are used in calculating the RPI for every month from February of that year to January of the following year.

Uses of price indices

- In the UK, the (annual) rate of inflation is the percentage increase in the value of the CPI (or the RPI) compared to one year earlier. (In M140, it will always be made clear whether you should use the CPI or the RPI.)
- To index-link any amount of money, the amount in question is multiplied by the same ratio as the change in the value of the price index.
- The purchasing power (in pence) of the pound at date A compared with date B is

$$\frac{\text{value of RPI at date } B}{\text{value of RPI at date } A} \times 100.$$

Unit 3 Earnings

Earnings ratios

In a context where men usually earn more than women, an earnings ratio will usually be less than one or, as a percentage, less than 100%. The nearer the earnings ratio is to 100%, the closer are the earnings of women to those of men.

The definitions of specific earnings ratios are as follows.

- At the mean:

$$\frac{\text{mean earnings of women}}{\text{mean earnings of men}}.$$

- At the median:

$$\frac{\text{median earnings of women}}{\text{median earnings of men}}.$$

- At the lower quartile:

$$\frac{\text{lower quartile earnings of women}}{\text{lower quartile earnings of men}}.$$

- At the upper quartile:

$$\frac{\text{upper quartile earnings of women}}{\text{upper quartile earnings of men}}.$$

- At the lowest decile:

$$\frac{\text{lowest decile earnings of women}}{\text{lowest decile earnings of men}}.$$

- At the highest decile:

$$\frac{\text{highest decile earnings of women}}{\text{highest decile earnings of men}}.$$

The Annual Survey of Hours and Earnings (ASHE)

The ASHE provides information on patterns of earnings and paid hours for employees within industries, occupations and regions of the UK. It covers people who are members of the Pay-As-You-Earn scheme.

The data collected include:

- total earnings for the pay-period including the specified week
- location of workplace
- occupation
- information concerning normal basic hours, overtime earnings and hours, bonus payments, pension contributions, and length of pay-period.

Skewness, the median and the mean

In right-skew data, the median is generally less than the mean.

In left-skew data, the median is generally greater than the mean.

Percentiles and deciles

In general the nth percentile is the number such that $n\%$ of the values in the batch fall below it.

Percentiles for which the corresponding percentage is a multiple of 10 are also called deciles. This is because they divide up the batch of data into tenths.

One tenth of the values are below the 10th percentile, one tenth are between the 10th and the 20th percentiles, and so on. In particular, the 10th percentile is called the lowest decile and the 90th percentile is called the highest decile.

Boxplots

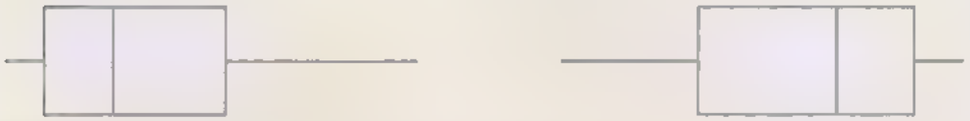
- A procedure for drawing a boxplot is as follows.
    1. The scale for the boxplot must run at least from the minimum to the maximum value in the batch. In M140, the boxplot is drawn so that the scale is horizontal.
    2. The 'box' of the boxplot runs from the lower quartile to the upper quartile. Within the box there is a line showing the position of the median.
    3. The 'whiskers' of the boxplot are lines, drawn parallel to the scale, that run from the lower quartile to the lower adjacent value, and from the upper quartile to the upper adjacent value. The lower/upper adjacent value is the furthest data value that is within one and a half times the IQR (interquartile range) of the lower/upper quartile.
    4. Any individual data values that are not covered by the box or the whiskers are plotted separately (in line with the whiskers). They are potential outliers.

- The shape of a batch of data can be determined from a boxplot.

  In a symmetric batch, the part of the boxplot to the right of the median is a mirror image of the part to the left of the median.

  In a right-skew batch, the right-hand part of the box (above the median) is longer than the left-hand part of the box (below the median). Also the right-hand whisker is longer than the left-hand whisker.

  In a left-skew batch, the left-hand part of the box (below the median) is longer than the right-hand part of the box (above the median). Also the left-hand whisker is longer than the right-hand whisker.



Right-skew Left-skew

Skewness in boxplots

- In a decile boxplot, the whiskers extend out as far as the lowest and highest deciles. Arrowheads are used at the ends of the whiskers.

Deviations, variance and standard deviation

- For each data value in a batch there is a deviation of the data value from the mean, or just deviation for short. When the batch mean is $\overline{x}$, the deviation for a data value x is $x - \overline{x}$.
- Deviations measure how far the data values in a batch are from the batch mean.
  - If a data value is exactly equal to the mean, then the deviation will be zero.
  - If a data value is close to the mean, then the deviation will be a small number (near zero).
  - If a data value is a long way above the mean, then the deviation will be large and positive.
  - If a data value is a long way below the mean, then the deviation will be large and negative.

  Thus if a batch has a large spread, its data values will tend to be a relatively long way from the mean, so the deviations will tend to be large in size. (Large negative numbers and large positive numbers are both large in size.)
- The quantity obtained, for a batch of size n, by calculating the sum of squared deviations and dividing by $n - 1$, is called the variance of the batch. It is a measure of spread.
- The standard deviation is the square root of the variance. It has the same units as the original data. It is also a measure of spread.

- The standard deviation can be calculated in the following ways.
  - Method 1: Calculate the mean $\bar{x}$, subtract it from each data value and hence work out $\sum(x-\bar{x})^2$.
  - Method 2: Calculate the sum of the data values, $\sum x$, and the sum of the squares of the data values, $\sum x^2$, and hence work out $\sum(x-\bar{x})^2 = \sum x^2 - \dfrac{(\sum x)^2}{n}$.

  Method 2 is easier to use when doing the calculations by hand.

  By either method,

  Variance: $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$.

  Standard deviation: $s = \sqrt{\text{variance}}$.

Grouped data

Data which are split into groups with the number of observations in each group recorded are called grouped data. Frequencies are the numbers that tell us how frequent the corresponding data values are.

A method for calculating the mean and standard deviation from grouped data by hand is as follows.

Denoting the data values by x and the corresponding frequencies by f:

1. Construct a table which includes the following columns: x, the data values; f, the frequencies; xf; and x^2f. Using this table, calculate the batch size, $n = \sum f$, the sum of the data values, $\sum xf$, and the sum of the squares of the data values, $\sum x^2 f$.
2. Calculate the mean as $\bar{x} = \dfrac{\sum xf}{n}$.
3. Calculate the sum of the squares of the deviations as
$$\sum(x-\bar{x})^2 f = \sum x^2 f - \frac{(\sum xf)^2}{n}.$$
4. Divide the result of step 3 by $n-1$, giving
$$\text{variance} = \frac{\sum(x-\bar{x})^2 f}{n-1}.$$
5. Calculate the standard deviation as $s = \sqrt{\text{variance}}$.

Summary measures

Summary measures are used to summarise important features of a batch, such as its location and spread. They include the mean, median, range, interquartile range, standard deviation and variance.

Factors which help determine which measure to use include the following.

- Consistency. Summary measures tend to be used in pairs such as: the mean and standard deviation; the median and interquartile range.

- Purpose. The same measure tends to used throughout an analysis, and a particular measure might be required in a later stage of the analysis.
- Resistance. The median and interquartile range are resistant measures, whereas the mean and standard deviation are not.

Real earnings for month A compared with one year earlier

The real earnings for month A compared with one year earlier is defined as:

$$\frac{\text{AWE index for month } A}{\text{AWE index for one year earlier}} \times \frac{\text{CPI for one year earlier}}{\text{CPI for month } A}$$

Unit 4 Surveys

Simple random sampling

Simple random sampling is a method of selecting a sample in which the possible samples of a given size, n, consist of all possible selections of n different individuals from the population. The sample to be used is then chosen in such a way that every possible sample is equally likely to be selected.

A way of doing this is to choose the sample members one at a time in such a way that:

- at each selection, every member of the target population is equally likely to be selected
- the selection of a particular member of the target population has no effect on the other selections, beyond the requirement that the same individual cannot appear more than once in the sample.

By numbering every member of the target population, random numbers can be used to select the sample. The random numbers must include all the labels for the population. Random numbers that correspond to labels which are not used, or correspond to a person already selected, are ignored.

Systematic random sampling

Systematic random sampling provides a quick and easy method of selecting a sample when the population is given in a list.

A 1 pth sample (p is the sampling interval) is taken using the following procedure.

1. Decide where to start by randomly choosing a label from the first p labels. This label is the random start.
2. Select the remaining individuals from the population by systematically selecting every pth label.

Systematic random sampling is likely to do better than simple random sampling when the list is first ordered by some strata (such as occupation and gender).

Systematic random sampling is likely to do worse than simple random sampling when the members from each of the strata appear in a systematic way in the list (such as male, female, male, female, ... in a list of married couples).

Sampling distributions

- The median obtained from a sample of size n is the median response or the sample median.
- The distribution of the median responses from all possible samples is the distribution of the median response of the sample, or distribution of the sample median for short.
- As the sample size increases, the distribution of the sample median becomes more clustered around the population median. This indicates that as the sample size increases, the median response obtained from any particular sample is more likely to be close to the population median.

Stratified sampling

- If a population is stratified, a stratified sample might then be chosen by selecting approximately the same proportion of individuals from each stratum. Such a stratified sample will be representative of the population with respect to the sizes of these strata. However, a stratified sample need not be chosen in this way, and often further knowledge about the population or the purpose of sampling will suggest better methods of selecting individuals from the strata.
- When approximately the same proportion of individuals are to be selected from each stratum,

$$\text{stratum subsample size} \simeq \frac{\text{sample size} \times \text{stratum size}}{\text{total population size}}$$

  (Any subsample size less than one would be set equal to one.)
- To be used in sampling, strata in the population must be
  - exhaustive: every member of the population must belong to a stratum
  - mutually exclusive: no member of the population can belong to more than one stratum
  - relevant to the subject under investigation: within each stratum, individuals should as far as possible be similar with respect to this subject
  - known for all population members before the sample is chosen: otherwise a list of the individuals in a stratum from which to choose the subsample would not be available.

Cluster sampling

Cluster sampling is a form of sampling that restricts the costs of surveys by restricting the sample to a limited number of geographical areas.

Cluster sampling works as follows.

- Find suitable geographical areas.
- Choose, preferably using random methods, a limited number of these geographical areas.
- For each of these chosen geographical areas choose a subsample from those members of the population in that area. For example, draw a simple random sample from each of these clusters. The clusters may differ in their sizes, and the sizes of the subsamples drawn from them should vary correspondingly: subsample approximately the same pre-specified proportion of each cluster.

Comparison of stratified and cluster sampling

Stratified sampling	Cluster sampling
• Each stratum focuses on one section of the population, such as those of a specified gender in a particular age group.	• Each cluster should be, as far as possible, a representative cross-section of the whole population.
• Every member of the population must be in one and only one stratum.	• Every member of the population must be in one and only one cluster.
• A stratified sample includes members of every stratum.	• A cluster sample excludes all the members of some (usually most) of the clusters.
• Stratified sampling decreases sampling error compared to a simple random sample of the same size (i.e. it is more efficient) but slightly increases costs.	• Cluster sampling often decreases costs but usually increases sampling error compared to a simple random sample of the same size (i.e. it is less efficient).

Quota sampling

Quota sampling is a procedure that is used frequently for market research surveys and opinion polls. First the sample size is determined (usually by consideration of costs), and then each interviewer is allocated a quota of interviews to achieve. The interviewers are then sent out to contact suitable respondents at selected sites in selected towns.

Quota sampling is economical because it produces quick results. However, a quota sample is not a random sample: the selection of individuals is haphazard rather than random. This means that it is usually difficult to give a numerical estimate for how unrepresentative the results are likely to be.

Errors in surveys

There are three types of error that can occur in surveys:

1. Sampling error. This is variability due to sampling.
2. Bias. Error introduced by using a poor sampling scheme.

3. Non-sampling errors. Errors which can arise from a variety of causes, such as: errors in recording responses or in transferring them to a computer; failure to contact individuals who are supposed to be included in a sample; or refusal of people to cooperate with the interviewer.

Sampling from the electoral register

One sampling frame that has commonly been used in the UK for surveys of individual adults and of households is the register of electors.

The electoral register lists all electors, and an edited version is available to buy. The full register contains almost all adults who are eligible to vote, as the registration of eligible voters is compulsory in the UK.

However, the electoral register does not contain many non-EU citizens nor people aged under 17. (People can be registered to vote from age 17, though their registration is not activated until they reach their 18th birthday.) The edited register does not include anybody who has chosen not to be included in the edited version. Also the electoral register is out-of-date even when it is first published, because compiling a relatively complete list of a large human population is time-consuming.

Unit 5 Relationships

Explanatory and response variables

A response variable is the variable that is being explained or whose value depends on other variables. It is also the variable to be predicted if predictions are to be made.

An explanatory variable is the variable that is doing the explaining or is the variable on which the response variable depends.

By convention, on a scatterplot the explanatory variable is put on the x-axis and the response variable is put on the y-axis.

Positive and negative relationships

On a scatterplot, variables are said to be positively related if low values of x are associated with low values of y, and high values of x are associated with high values of y.

That is, if points tend to slope *upwards* from left to right, then the variables are *positively* related.

Variables are said to be negatively related if low values of x are associated with high values of y, and high values of x are associated with low values of y.

That is, if points tend to slope *downwards* from left to right, then the variables are *negatively* related.

Linear and non-linear relationships

A relationship is said to be linear if it can be summarised reasonably well by a straight line.

A relationship is said to be non-linear if it can be summarised reasonably well by a curve but not by a straight line.

Strong and weak relationships

A relationship is said to be strong when all the points on a scatterplot lie close to a line. A relationship is said to be weak when all the points only loosely follow a line.

There is said to be no relationship between two variables when knowledge of the value of the explanatory variable does not provide information about the value of the response variable.

Interpretation checklist for scatterplots

- Is the relationship positive, negative or neither?
- Is the relationship linear or non-linear?
- Is the relationship strong or weak?
- Are there any outliers?

Fitting lines

When summarising data on a scatterplot, the simplest adequate curve should be chosen. In many cases this amounts to choosing an appropriate straight line.

The process of choosing a straight line to draw is often called fitting a line to the data. There are many different ways to do this. One method is simply to draw in the line that appears to give a good representation of the pattern in the data.

Least squares is another method used to fit lines to data. This method has the advantage that it is a formal method that can be carried out by computers.

DFR equation and residuals

- The DFR equation splits an observed 'Data' value into a 'Fit' value and a 'Residual' value. These are linked in the following way.

  $$\text{Data} = \text{Fit} + \text{Residual}$$

  $$\text{Residual} = \text{Data} - \text{Fit}.$$

- On a scatterplot for each point, 'Data' is the position of that point up the y-axis. 'Fit' is the vertical position of the line for the value of the explanatory variable. The 'Residual' is a measure of the vertical distance from the data point to the fitted line.
- If the residuals for a fit line show a pattern that relates to the explanatory variable, the fit line does not provide an adequate explanation of all the pattern in the data, and we should look for a better relationship.

The least squares regression line

The least squares regression line is the line for which the sum of the squares of the residuals is minimised.

The least squares regression line always goes through the point $(\bar{x}, \bar{y})$, and hence the sum of the residuals is always zero.

For a set of n data points (x, y), the least squares regression line $y = a + bx$ can be calculated in the following way.

1. Calculate $\sum x$, $\sum y$, $\sum x^2$ and $\sum xy$.
2. Calculate the means of x and y:

$$\bar{x} = \frac{\sum x}{n} \quad \text{and} \quad \bar{y} = \frac{\sum y}{n}.$$

3. Calculate the sum of the squared deviations of the x-values

$$\sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n},$$

and the sum of the products of the deviations

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}.$$

4. The slope b is given by

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}.$$

5. The intercept a is given by

$$a = \bar{y} - b\bar{x}.$$

Using fit lines

The fitted value $y = a + bx$ is an estimate of the average value of the response variable Y that occurs when the explanatory variable takes the value $X = x$.

The residual associated with each point is given by the following formula.

$$\begin{aligned} \text{Residual} &= \text{Data} - \text{Fit} \\ &= y - (a + bx). \end{aligned}$$

Prediction should only be done from explanatory variable to response. Predictions only apply to populations from which the original data were taken and are only valid for the range of values of x, the explanatory variable, represented in the original sample.

Unit 6 Truancy

Statistical inference

This is a process of inferring back from the sample to the population. That is, making inferences about a population on the basis of a sample of data taken from that population.

Somewhat paradoxically, the best approach to making inferences about populations from a sample is to consider what samples are likely to be obtained from a population.

Truancy rate

One of the statistics on schools that the government publishes is the unauthorised absence rate. We use this in M140 as our truancy rate.

- A pupil's truancy rate is the proportion of school half-days that the pupil was absent without authorisation.
- A school's truancy rate is the average truancy rate of its pupils.

Probability

- In general, the probability of an event is the proportion of times that event is likely to occur. In particular, if E stands for the event of selecting a person or object with some particular property from a population using random sampling, then the probability of E is given by

$$P(E) = \frac{\text{number in population with particular property}}{\text{total number in population}}.$$

  - If an event is impossible, then its probability is 0.
  - If an event is certain, then its probability is 1.
  - Any event which is uncertain but not impossible has a probability that lies between 0 and 1.

- Two events are said to be mutually exclusive if they cannot occur at the same time. More generally, any number of events are said to be mutually exclusive if no two of them can occur at the same time.
- Two events are said to be statistically independent if the occurrence of one has no effect on the likelihood of occurrence of the other.
- For any two mutually exclusive events A and B,

$$P(A \text{ or } B) = P(A) + P(B).$$

  This rule extends to more than two events when they are all mutually exclusive. For example, if A, B and C are mutually exclusive events, then

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C).$$

- If A and B are statistically independent events, then the probability that A and B both occur is given by

  $$P(A \text{ and } B) = P(A) \times P(B).$$

  This rule extends to more than two events. For example, if A, B and C are statistically independent events, then the probability that they all occur is given by

  $$P(A \text{ and } B \text{ and } C) = P(A) \times P(B) \times P(C).$$

- For any event E,

  $$P(E \text{ does not occur}) = 1 - P(E \text{ does occur}).$$

Choosing a sample from a set of objects

Suppose there are n objects to choose from.

- The number of ways of choosing x objects in a *specified order* is

  $$n \times (n-1) \times \cdots \times (n-x+1).$$

  (There are x terms in $n \times (n-1) \times \cdots \times (n-x+1)$.)

- The number of ways of choosing x objects *if the order does not matter* is

  $$^nC_x = \frac{\text{number of choices of } x \text{ objects if order does matter}}{\text{number of ways in which } x \text{ objects can be ordered}}$$
  $$= \frac{n \times (n-1) \times \cdots \times (n-x+1)}{x \times (x-1) \times \cdots \times 1}.$$

  (There are x terms in both $n \times (n-1) \times \cdots \times (n-x+1)$ and $x \times (x-1) \times \cdots \times 1$.)

  It can be seen directly from the definition that for any value of n, $^nC_0 = 1$ and $^nC_n = 1$.

Probability of the number of observations above a median

Suppose we take a random sample of size n. Then

$$^nC_x \times \left(\frac{1}{2}\right)^n$$

is the probability that exactly x of these observations are greater than the population median. This formula is a special case of a probability distribution known as the binomial distribution.

Steps in a hypothesis test

1. Make a statement about the population of interest (e.g. *the median truancy rate of large schools in the East of England is 0.98%*) This is the hypothesis we wish to test.
2. Under the assumption that the hypothesis is true, determine the probability distribution for all possible values of some sample statistic (e.g. determine the probability distribution for the number of large schools, out of 12, that will have a truancy rate above 0.98%).
3. Now take the sample and ascertain *how unlikely* the observed value of the sample statistic is, on the basis of (1) and (2) (e.g. are we very unlikely to get a sample statistic as large as 9, that is, a sample of 12 schools in which nine or more schools have a truancy rate above 0.98%?).
4. If the sample turns out to have a very unlikely value, then either:
   - a very unusual event has happened, or
   - the hypothesis suggested in step 1 is incorrect, in which case the sample has provided evidence, albeit in a negative way, that adds something to our beliefs about the population.

Critical region, critical values and significance levels

- The significance level is the smallest p-value, expressed as a percentage, at which the hypothesis will not be rejected. If a sample is selected whose values are one of the 5% most extreme outcomes that might occur if the hypothesis is true, then we reject the hypothesis at the 5% significance level.
- The critical region at the 5% significance level is chosen so that the combined probability of the outcomes falling in the region is 0.05 or just less than that.
- The critical value at the 5% significance level can be used to determine whether or not an outcome is in the critical region.

Procedure: the sign test

1. State the hypothesis that the population median is M.
2. Count the number of values in the sample that are larger than M (denoted by [+]s) and the number of values that are smaller than M (denoted by [−]s). The smaller of these two values is the test statistic.
3. Use Table 8 (at the end of Subsection 4.1, and replicated at the end of this Handbook) to write down the critical value at the 5% significance level corresponding to the size of the sample.
4. Compare the test statistic with the critical value. If it is less than or equal to the critical value, then the hypothesis is rejected at the 5% significance level. If the test statistic is greater than the critical value, then the hypothesis is not rejected.

For a sample of size n containing m ties (that is, m of the sample values are equal to the assumed median), discard the m ties and treat the sample as one of size $(n - m)$.

Procedure: obtaining p-values

To obtain the p-value (significance probability) for a hypothesis test, work through the following steps.

1. Assume the hypothesis is true.
2. Consider all the possible outcomes and divide these into two sets:

   Set A contains those outcomes that are as extreme or more extreme than the outcome that actually occurred.

   Set B contains those outcomes that are more likely than the outcome that actually occurred.
3. Calculate the probability that a random outcome would be from Set A. This probability is the p-value.

Interpretation of p-values

The following table gives the interpretation of p-values.

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

Unit 7 Factors affecting reading

The normal distribution

- Normal distributions are important both as (approximate) population distributions, in some cases, and as (approximate) sampling distributions, in many more cases.
- The normal distribution is 'bell-shaped'. It has location specified by the population mean μ and spread specified by the population standard deviation σ.
- The normal distribution has its mode at μ.
- Almost the whole of the normal distribution is contained between $\mu - 3\sigma$ and $\mu + 3\sigma$.
- The standard normal distribution is the particular normal distribution that has mean $\mu = 0$ and standard deviation $\sigma = 1$.
- The standard normal distribution is important because we can think of all normal distributions in terms of it. If a variable x has a normal distribution with mean μ and standard deviation σ, then the variable $z = (x - \mu)/\sigma$ has the standard normal distribution. The variable z corresponds to the number of standard deviations by which the variable x differs from its mean.

Approximate sampling distribution of the mean

For most practical purposes, whatever the shape of the original population distribution, the sampling distribution of the mean for samples of size n will be approximately normal if n is a reasonable size (say $n \geq 25$). This important result is often called the central limit theorem.

The mean of the sampling distribution is equal to μ, the population mean.

The standard deviation of the sampling distribution is called the standard error of the mean, and is given by $\text{SE} = \sigma/\sqrt{n}$, where n is the sample size and σ is the population standard deviation.

The one-sample z-test

For a one-sample z-test, the hypotheses are concerned with the mean, μ, of the population from which the sample is selected.

The null hypothesis is H_0: $\mu = A$, and the alternative hypothesis is H_1: $\mu \neq A$. (This is a two-sided alternative hypothesis; one-sided alternative hypotheses will be discussed in Unit 10.)

The information you need to know for a one-sample z-test is:

- the hypothesised population mean (A) under the null hypothesis
- the sample mean ($\overline{x}$)
- the sample size (n)
- the population standard deviation (σ), or a good estimate of σ.

As a rough guide you can assume that, whatever the population distribution, for sample sizes of at least 25, the one-sample z-test is applicable.

Procedure: the one-sample z-test

1. Set up the null and alternative hypotheses,

   H_0: $\mu = A$
   H_1: $\mu \neq A$,

   where μ is the population mean.

2. Calculate the test statistic, z:

   - If the population standard deviation (σ) is known,

     $$z = \frac{\overline{x} - A}{\text{SE}}, \quad \text{where SE} = \frac{\sigma}{\sqrt{n}}.$$

   - If σ is unknown but the sample size (n) is 25 or more,

     $$z = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

     Here $\overline{x}$ is the sample mean and s is the standard deviation of the sample. SE is the standard error of the mean and ESE is the estimated standard error.

3. Compare z with the appropriate critical values, which are 1.96 and -1.96 at the 5% significance level and 2.58 and -2.58 at the 1% significance level.
   - If $z \geq 2.58$ or $z \leq -2.58$, then H_0 is rejected at the 1% significance level.
   - If $1.96 \leq z < 2.58$ or $-2.58 < z \leq -1.96$, then H_0 is rejected at the 5% significance level but not at the 1% significance level.
   - If $-1.96 < z < 1.96$, then H_0 is not rejected at the 5% significance level.
4. State the conclusions that can be drawn from the test.

Approximate sampling distribution of the difference between two means

Suppose samples of size n_A and n_B are taken from two different populations A and B. If n_A and n_B are large, no matter what shape the population distributions, the sampling distribution of the difference between two means based on these samples will in practice be approximately normal.

The mean of the sampling distribution is equal to $\mu_A - \mu_B$, the difference between the population means.

The standard deviation of the sampling distribution is called the standard error of the difference between two means, and is given by

$$\text{SE} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}},$$

where σ_A and σ_B are the population standard deviations.

The two-sample z-test

The two-sample z-test is used to analyse the difference in means between two populations.

In general, call the two groups A and B. The null and alternative hypotheses are:

H_0: $\mu_A = \mu_B$
H_1: $\mu_A \neq \mu_B$,

where μ_A and μ_B are the means of populations A and B, respectively.

The information you need to know for a two-sample z-test is:

- the sample means ($\bar{x}_A$ and $\bar{x}_B$)
- the sample sizes (n_A and n_B)
- the population standard deviations (σ_A and σ_B), or good estimates of them (s_A and s_B).

As a rough guide you can assume that, whatever the population distribution, when both n_A and n_B are at least 25, the two-sample z-test is applicable.

Procedure: the two-sample z-test

1. Set up the null and alternative hypotheses,

   H_0: $\mu_A = \mu_B$
   H_1: $\mu_A \neq \mu_B$,

   where μ_A and μ_B are the means of populations A and B, respectively.

2. Calculate the test statistic

   $$z = \frac{\bar{x}_A - \bar{x}_B}{\text{ESE}},$$

   where the estimated standard error of $\bar{x}_A - \bar{x}_B$ is

   $$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

   Here, n_A and n_B are the sample sizes of random samples from populations A and B respectively, $\bar{x}_A$ and $\bar{x}_B$ are the sample means, and s_A and s_B are the sample standard deviations.

3. Compare z with the appropriate critical values, which are 1.96 and -1.96 at the 5% significance level, and 2.58 and -2.58 at the 1% significance level.

   - If $z \geq 2.58$ or $z \leq -2.58$, then H_0 is rejected at the 1% significance level.
   - If $1.96 \leq z < 2.58$ or $-2.58 < z \leq -1.96$, then H_0 is rejected at the 5% significance level but not at the 1% significance level.
   - If $-1.96 < z < 1.96$, then H_0 is not rejected at the 5% significance level.

4. State the conclusions that can be drawn from the test.

Unit 8 Teaching how to read

Contingency tables

A contingency table is a table which meets the following three conditions:

- the row variable and the column variable are both categorical
- the categories for both variables are mutually exclusive
- the entry in each cell of the table is a count.

In a contingency table, the row and column totals are known as the marginal totals. The dimension (or size) of a contingency table with r rows and c columns (excluding totals) is $r \times c$.

Probabilities can be obtained from contingency tables by calculating proportions.

Joint probabilities, conditional probabilities and statistical independence

Let A and B denote two events, mutually exclusive or not.

- The joint probability of A and B, denoted $P(A \text{ and } B)$, is the probability that both A and B occur.
- The conditional probability of A given B, denoted $P(A|B)$, is the probability that A occurs, given that B occurs.
- Joint and conditional probabilities are linked by the following relationships:

  $P(A \text{ and } B) = P(A) \times P(B|A) = P(B) \times P(A|B)$.

  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.
- A and B are statistically independent if the occurrence of one has no influence on the chance of occurrence of the other.

  Moreover, they are independent exactly when

  $$P(A|B) = P(A)$$

  or (equivalently)

  $$P(B|A) = P(B).$$
- Two variables are said to be independent if the value taken by either variable does not influence the value taken by the other variable.

The χ^2 test for contingency tables

The χ^2 test for contingency tables is a hypothesis test to investigate whether two variables given in a contingency table are independent.

The χ^2 test for contingency tables should be used only when all Expected values are greater than or equal to 5.

Procedure: the χ^2 test for contingency tables

The procedure for the χ^2 test for contingency tables is as follows.

1. Set up the null and alternative hypotheses in terms of the independence or otherwise of the row and column variables.
2. From the Observed table, calculate the Expected, Residual and χ^2 contribution tables. That is, for each cell in a contingency table calculate
   - the Expected value (E) under the null hypothesis of independence.

     $$\frac{\text{row total} \times \text{column total}}{\text{overall total}}$$
   - the Residual value, $O - E$
   - the χ^2 contribution, $\dfrac{(O-E)^2}{E}$.
3. Calculate the χ^2 test statistic by summing all the χ^2 contributions.

4. Find the degrees of freedom: $(r-1) \times (c-1)$. Look up the critical values at the 5% and 1% significance levels (CV5 and CV1), using Table 28 of Unit 8 (Subsection 4.4) or the extended version at the end of this Handbook.
5. Compare χ^2 with CV5 and CV1.
   - If $\chi^2 \geq$ CV1, then H_0 is rejected at the 1% significance level.
   - If CV5 $\leq \chi^2 <$ CV1, then H_0 is rejected at the 5% significance level but not at the 1% significance level.
   - If $\chi^2 <$ CV5, then H_0 is not rejected at the 5% significance level.
6. State your conclusion in non-mathematical terms.

Type 1 and type 2 errors

- The use of a hypothesis test always carries with it the possibility of error.
- A type 1 error occurs when the null hypothesis is rejected even though it is true. A type 2 error occurs when the null hypothesis is not rejected even though it is actually false.

  These definitions are summarised in the following table.

	H_0 true	H_0 false
H_0 not rejected	Correct	Type 2 error
H_0 rejected	Type 1 error	Correct

- Type 1 and type 2 errors are related for a given dataset as the smaller the probability of making a type 1 error, the larger the probability of making a type 2 error. The accepted procedure is to fix the probability of making a type 1 error. The probability of making a type 1 error is 0.05 if we use a 5% significance level and 0.01 if we use a 1% significance level.

Unit 9 Comparing schools

The correlation coefficient

The correlation coefficient is a number which summarises the strength of relationship between two variables.

- The correlation coefficient does not depend on the scales of the axes. It only reflects the pattern of the points.
- The correlation coefficient does not depend on which variable is plotted on the vertical axis and which is plotted on the horizontal axis.
- A correlation coefficient close to zero does not imply that there is no relationship, merely that there is not a linear relationship.
- Removing an influential point reduces the amount of correlation (i.e. moves the correlation coefficient closer to zero), whereas removing an outlier usually increases it (moves the coefficient towards +1 or −1).
- Correlation is not causation.

Calculating the correlation coefficient

Given a batch of n linked data pairs, (x, y), the correlation coefficient (r) is obtained as follows:

1. Calculate $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ and $\sum xy$.
2. Calculate

$$\sum (x-\bar{x})^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2,$$
$$\sum (y-\bar{y})^2 = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2.$$
$$\sum (x-\bar{x})(y-\bar{y}) = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right).$$

3. Use the values from step 2 to calculate

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \times \sum (y-\bar{y})^2}}.$$

Confidence intervals

- A 95% confidence interval for μ includes all values of A for which we cannot reject H_0: $\mu = A$ at the 5% significance level.

  A 99% confidence interval for μ includes all values of A for which we cannot reject H_0: $\mu = A$ at the 1% significance level.

- About 95% of the possible random samples we could select will give rise to a 95% confidence interval that does include the population mean. About 2.5% will give intervals that are completely below the population mean, and about 2.5% will give intervals completely above it. Thus, about 5% of possible random samples that might be selected will give rise to a 95% confidence interval that does not include the population mean.

  So if you say that a 95% confidence interval includes the population mean, you will be right 95% of the time; that is, you can be 95% confident that your statement is correct.

- If the confidence interval does include the hypothesised population mean, we do not reject the hypothesis. If the confidence interval does not include the hypothesised population mean, then we do reject the hypothesis. In particular:
  - If the 95% confidence interval does not include the hypothesised population mean, then we reject the hypothesis at the 5% significance level.
  - If the 99% confidence interval does not include the hypothesised population mean, then we reject the hypothesis at the 1% significance level.

Calculating 95% and 99% confidence intervals for a population mean

Suppose the sample size is n, the sample mean is $\bar{x}$ and the sample standard deviation is s.

- Calculate the estimated standard error: $\text{ESE} = s/\sqrt{n}$.
- The 95% confidence interval for the population mean is

  $$(\bar{x} - 1.96\,\text{ESE},\ \bar{x} + 1.96\,\text{ESE}).$$

- The 99% confidence interval for the population mean is

  $$(\bar{x} - 2.58\,\text{ESE},\ \bar{x} + 2.58\,\text{ESE}).$$

As with the z-test, these formulas should only be used if the sample size is at least 25.

Calculating 95% and 99% confidence intervals for $\mu_A - \mu_B$, the difference between two population means

Suppose the two sample sizes are n_A and n_B, the sample means are $\bar{x}_A$ and $\bar{x}_B$, and the sample standard deviations are s_A and s_B.

- Calculate the estimated standard error:

  $$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

- The 95% confidence interval for $\mu_A - \mu_B$ is

  $$(\bar{x}_A - \bar{x}_B - 1.96\,\text{ESE},\ \bar{x}_A - \bar{x}_B + 1.96\,\text{ESE}).$$

- The 99% confidence interval for $\mu_A - \mu_B$ is

  $$(\bar{x}_A - \bar{x}_B - 2.58\,\text{ESE},\ \bar{x}_A - \bar{x}_B + 2.58\,\text{ESE}).$$

These formulas should only be used if the sample sizes are at least 25.

Confidence intervals for the mean response

A confidence interval for the mean response is an interval estimate for the vertical position of a least squares regression line for a particular value of the explanatory variable, x.

- Confidence intervals for the mean response are at their narrowest when x is the sample mean, and get steadily wider each side.
- The least squares regression line based on the sample data is always in the middle of the interval. Hence the point estimate for the predicted value is always in the middle of the confidence interval for the mean response.

Prediction intervals

Prediction intervals reflect the random variation of individual values around the population regression line as well as uncertainty about the actual position of that line. If we have a very large sample, then we will have a good idea about the position of the population regression line – most of the uncertainty in predicting the value of an individual will stem from the scatter of individual values about the regression line.

A prediction interval has the following properties:

- it is centered around the predicted value $a + bx$
- it is narrowest when x is the sample mean, and steadily widens away from this point
- it gets wider as the scatter around the line increases
- it is always wider than the corresponding confidence interval for the mean response.

Unit 10 Experiments

What is an experiment?

Good experiments share two important features:

- They are recorded in detail so that they can be critically evaluated and repeated.
- They produce measurements or observations designed to answer specific questions.

At least three kinds of experiment can be recognised: they are distinguished by the kind of questions they attempt to answer. These are:

- Exploratory (Baconian) experiments
- Measurement experiments
- Hypothesis-testing (hypothetico-deductive) experiments.

The basis of statistical hypothesis tests

When we carry out a statistical hypothesis test, we base our calculations on the assumption that the null hypothesis is correct. We ask:

> *What is the probability of obtaining a result at least as extreme as that which we have obtained, if we assume that the null hypothesis is true?*

If this probability is too low, then we reject the null hypothesis in favour of the alternative.

The two-sample t-test (assuming a common population variance)

Like the two-sample z-test, the two-sample t-test is used to analyse the difference in means between two populations.

For the two-sample t-test:

- The data must be numerical measurements (such as length, weight, time) that form two unrelated samples.
- It is assumed that each sample is selected from a population whose distribution is normal.
- It is assumed that the standard deviations of the two populations are equal or, equivalently, that the population variances are equal.

For the two-sample t-test, you need to know:

- the sample means ($\overline{x}_A$ and $\overline{x}_B$)
- the sample sizes (n_A and n_B)

- the sample standard deviations (s_A and s_B), or the pooled standard deviation (s_p), or the corresponding variances.

If the sample variances (s_A^2 and s_B^2) differ by a factor of less than three, assume that there is a common population variance, or that, if the population variances differ, the difference is not large enough to invalidate the t-test.

Note that the two-sample t-test is suitable when the sample sizes, n_A and n_B, are not large.

Procedure: the (two-sided) two-sample t-test (assuming a common population variance)

1. Set up the null and alternative hypotheses

   H_0: $\mu_A = \mu_B$
   H_1: $\mu_A \neq \mu_B$,

   where μ_A and μ_B are means of the populations A and B respectively.
2. Calculate the sample means, $\bar{x}_A$ and $\bar{x}_B$, of the two samples and the sample variances, s_A^2 and s_B^2. (s_A and s_B are the sample standard deviations.)
3. Check that the assumption of equal population variances is reasonable, or that the assumption is not seriously violated.
4. Calculate a pooled estimate s_p^2 of the common population variance:

   $$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2},$$

   where n_A and n_B are the two sample sizes.
5. Calculate the test statistic:

   $$t = \frac{\bar{x}_A - \bar{x}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}.$$

6. Calculate the degrees of freedom: $n_A + n_B - 2$. Look up the critical value, t_c, which is based on the t distribution.
7. Compare the test statistic t with t_c.
   - If $t \geq t_c$ or if $t \leq -t_c$, then H_0 is rejected at the 5% significance level.
   - If $-t_c < t < t_c$, then H_0 is not rejected at the 5% significance level.
8. State the conclusions that can be drawn from the test.

The one-sample t-test

Similarly to the one-sample z-test, the hypotheses for the one-sample t-test are concerned with the mean μ of the population from which the sample is selected.

The null hypothesis is H_0: $\mu = A$, and the (two-sided) alternative hypothesis is H_1: $\mu \neq A$.

The information you need to know for a one-sample t-test is:

- the hypothesised population mean (A) under the null hypothesis
- the sample mean ($\overline{x}$)
- the sample size (n)
- the sample standard deviation (s)

For the one-sample t-test, the sample size does not have to be large. However, it should be reasonable to assume that the population distribution is normal.

The t-test is said to be a more powerful test than the sign test because the t-test is better at identifying a null hypothesis that is false.

Procedure: the (two-sided) one-sample t-test

1. Set up the null and alternative hypotheses

   H_0: $\mu = A$
   H_1: $\mu \neq A$,

   where μ is the population mean.

2. Calculate the test statistic

   $$t = \frac{\overline{x} - A}{\text{ESE}},$$

   where the estimated standard error of $\overline{x}$ is

   $$\text{ESE} = \frac{s}{\sqrt{n}}.$$

   Here, $\overline{x}$ is the sample mean, s is the sample standard deviation and n is the sample size.

3. Calculate the degrees of freedom: $n - 1$. Look up the critical value, t_c, which is based on the t distribution.

4. Compare the test statistic t with t_c.

   - If $t \geq t_c$ or if $t \leq -t_c$, then H_0 is rejected at the 5% significance level.
   - If $-t_c < t < t_c$, then H_0 is not rejected at the 5% significance level.

5. State the conclusions that can be drawn from the test.

The matched-pairs t-test

In a matched-pairs experiment, items are paired in such a way that the factor of interest (but little else) differs between the two items that form a pair. The statistical analysis is then based on the differences between items within a pair.

If it can be assumed that this population of differences has a normal distribution, then apply the one-sample t-test with $A = 0$ and d instead of x in the formulas, to the sample of differences.

Procedure: the (two-sided) matched-pairs t-test

1. Calculate the differences between the two values in each pair.
2. The null and alternative hypotheses are

   H_0: $\mu_A = \mu_B$
   H_1: $\mu_A \neq \mu_B$,

   where μ_A and μ_B are the population means of the two populations involved.

   Replace these by the equivalent hypotheses

   H_0: $\mu_d = 0$
   H_1: $\mu_d \neq 0$,

   where μ_d is the population mean of the population of differences between the matched pairs.
3. Apply the (two-sided) one-sample t-test to the differences.

Confidence intervals for means

In general, the lower limit of the confidence interval is

point estimate − (z or t critical value) × ESE,

and the upper limit is

point estimate + (z or t critical value) × ESE.

where ESE is the estimated standard error of the point estimate.

- The 95% confidence interval for the population mean of a normally distributed population is given by

  $$\left(\overline{x} - t_c \frac{s}{\sqrt{n}},\ \overline{x} + t_c \frac{s}{\sqrt{n}}\right).$$

  where n, $\overline{x}$, t_c and s are as in the procedure for the (two-sided) one-sample t-test. Thus t_c is the critical value for the t distribution with $n - 1$ degrees of freedom used for a two-sided t-test.
- When the data are matched pairs and the differences are normally distributed, the 95% confidence interval for the difference between population means is given by

  $$\left(\overline{d} - t_c \frac{s}{\sqrt{n}},\ \overline{d} + t_c \frac{s}{\sqrt{n}}\right).$$

  where $\overline{d}$ is the mean of the differences in the sample and t_c is the critical value for the t distribution with $n - 1$ degrees of freedom used for a two-sided t-test.
- The 95% confidence interval for the difference between the population means of two unrelated normally distributed populations with equal standard deviations is given by

  $$\left((\overline{x}_A - \overline{x}_B) - t_c\, s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}},\ (\overline{x}_A - \overline{x}_B) + t_c\, s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}\right),$$

where t_c is the critical value for the t distribution with $n_A + n_B - 2$ degrees of freedom used for a two-sided t-test.

The assumption of equal population standard deviations holds acceptably well if the ratio of the larger sample variance to the smaller sample variance is less than 3. This condition should be checked before forming a confidence interval for the difference between two population means on the basis of two unrelated samples, unless the condition has already been checked in the course of a hypothesis test.

Procedure: the one-sided t-test for one sample or matched-pairs samples

1. Set up the null and alternative hypotheses

   H_0: $\mu = A$

   H_1: $\mu > A$ OR H_1: $\mu < A$.

   where μ is the population mean.

2. Calculate the test statistic:

   $$t = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

   Here, $\overline{x}$ is the sample mean (or sample mean of the differences), s is the sample standard deviation and n is the sample size.

   Note that this test statistic is the same as would be calculated for the two-sided one-sample/matched-pairs t-test.

3. Calculate the degrees of freedom: $n - 1$. Look up the critical value, t_c, for a one-sided t-test.

4. Compare the test statistic t with t_c.

   If the alternative hypothesis is H_1: $\mu > A$, then

   - H_0 is rejected at the 5% significance level if $t > t_c$
   - H_0 is not rejected at the 5% significance level if $t \leq t_c$.

   If the alternative hypothesis is H_1: $\mu < A$, then

   - H_0 is rejected at the 5% significance level if $t < -t_c$
   - H_0 is not rejected at the 5% significance level if $t \geq -t_c$.

5. State the conclusions that can be drawn from the test.

Procedure: the one-sided two-sample t-test (assuming a common population variance)

1. Set up the null and alternative hypotheses

   H_0: $\mu_A = \mu_B$

   H_1: $\mu_A > \mu_B$ OR H_1: $\mu_A < \mu_B$,

   where μ is the population mean.

2. Calculate the sample means, $\overline{x}_A$ and $\overline{x}_B$, of the two samples and the sample variances, s_A^2 and s_B^2. (s_A and s_B are the sample standard deviations.)

3. Check that the assumption of equal population variances is reasonable, or that the assumption is not seriously violated.
4. Calculate a pooled estimate s_p^2 of the common population variance:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}.$$

where n_A and n_B are the two sample sizes.

5. Calculate the test statistic:

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}.$$

Note that this test statistic is the same as would be calculated for the two-sided two-sample t-test.

6. Calculate the degrees of freedom: $n_A + n_B - 2$. Look up the critical value, t_c, for a one-sided t-test.
7. Compare the test statistic t with t_c.

   If the alternative hypothesis is H_1: $\mu_A > \mu_B$, then

   - H_0 is rejected at the 5% significance level if $t > t_c$.
   - H_0 is not rejected at the 5% significance level if $t \leq t_c$.

   If the alternative hypothesis is H_1: $\mu_A < \mu_B$, then

   - H_0 is rejected at the 5% significance level if $t < -t_c$
   - H_0 is not rejected at the 5% significance level if $t \geq -t_c$.

8. State the conclusions that can be drawn from the test.

Unit 11 Testing new drugs

Placebos

Treating a patient can quite commonly appear to have a therapeutic effect (or can *actually* have a therapeutic effect), even when the treatment contains no medication and should be ineffectual. This is called the placebo effect.

A dummy treatment that superficially resembles the treatment being tested but contains no active ingredient is called a placebo.

Placebo-controlled trial

In a placebo-controlled trial, there is a treatment group and a control group. People in the treatment group receive the treatment being tested, while those in the control group are given a placebo.

Blinding

A trial where neither patients nor doctors know which treatment is administered, is called a double-blind trial. A study in which the patient is blind but the doctor is not, or vice versa, is sometimes called a single-blind trial.

Clinical trials

Three types of clinical trial design were discussed in M140: crossover trials, matched-pairs trials and group-comparative trials.

- In a crossover trial, each person acts as his or her own control: during the course of the trial, each person crosses over from having one treatment to having the other, or vice versa.

  The crossover trial eliminates the variability that would arise from using different people in the experimental and control groups, but it cannot be used when the experimental treatment irreversibly alters a patient's condition, nor is it suited to short-lasting diseases.

- In a matched-pairs trial, each person in one group is matched as closely as possible with a person in the other group.

  The matched-pairs trial eliminates much of the variability that arises from using different people as experimental and control subjects, but it can be difficult to achieve a good match between the experimental and control groups.

- In a group-comparative trial, people are allocated randomly to two groups, usually in such a way that the two groups contain approximately the same number of people.

  The group-comparative trial does not eliminate the variability that arises from using different individuals in the experimental and control groups, but it is relatively easy to set up.

Data, design and tests

The links between the data collected, the design of a trial and the hypothesis test used to analyse the data, can be summarised as follows.

Type of data	Design	Test
Categorical		χ^2
Interval scale	Group comparative	two-sample z- or t-test
Interval scale	Matched pairs	matched-pairs z- or t-test
Interval scale	Crossover	matched-pairs z- or t-test

Patient-years

If a single patient takes a drug for one year, then that constitutes one patient-year of use of the drug.

If two patients take the drug for six months each, then the usage is half a patient-year each, which again comes to one patient-year. Assuming that the patients take the same dose of the drug each day, the amount of drug consumed by one patient taking it for a year is the same as the amount consumed by two patients over six months, so a patient-year corresponds to the use of a certain amount of the drug.

If twelve patients each take the drug for a month, or if 365 patients each take it for a day, the total usage is still one patient-year.

Phases of trials

- Phase 1: Early clinical pharmacology. In a phase 1 trial the drug is usually given in increasing doses to healthy volunteers so as to evaluate biological action and safety.
- Phase 2: Early clinical investigations. The studies in phase 2 are usually the first studies in which the drug is given to patients with the condition that the drug is designed to help.
- Phase 3: Comparative studies. In this phase, treatment with the new drug is compared with existing therapies in a wider range of contexts.
- Phase 4: Post-marketing studies. Phase 4 studies use larger samples of patients than can be obtained before marketing. They aim to obtain further evidence about the safety of the drug.

Unit 12 Review

Summarising data

In summarising data, the following is the order of priority.

1. If data must be summarised by just one number, then a number that represents the location of the data should be given (usually the median or mean).
2. If two numbers are to be used as the summary, then the second number should indicate the spread of the data (usually the interquartile range, standard deviation or variance).
3. Additional numbers would describe the shape of the data, notably any skewness, and identify the largest and smallest data along with any numbers that are extreme relative to the main body of the data.

Graphical summaries can convey a lot of information in an accessible way.

Designing experiments

When designing experiments, you should aim to:

- give fair comparison of the different treatments being examined
- remove or reduce potential sources of bias
- try to reduce random variation
- gather as much data as you reasonably can.

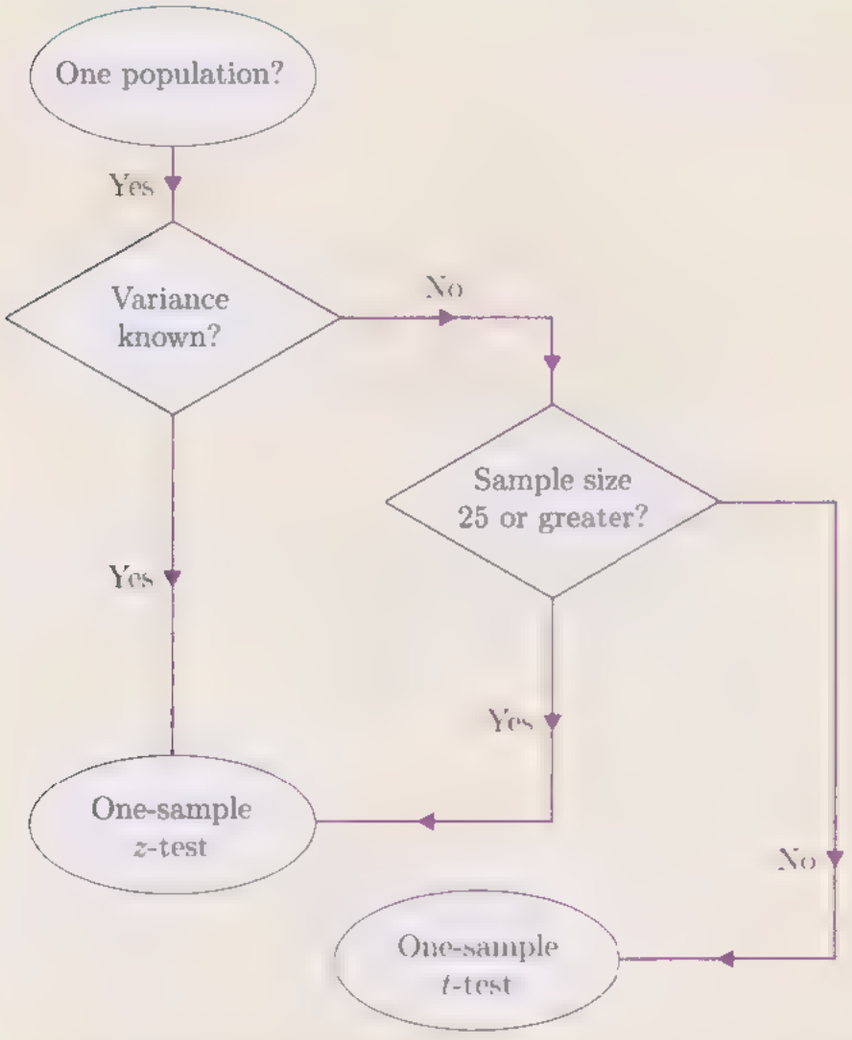
Binomial distribution

Suppose the result of a trial is success or failure (no other possibilities). Suppose also that the probability of success (p) is the same in each trial. Let $q = 1 - p$ and let x denote the number of successes in n trials. If trials are independent of each other, then

$$P(x) = {}^nC_x \times p^x \times q^{n-x}.$$

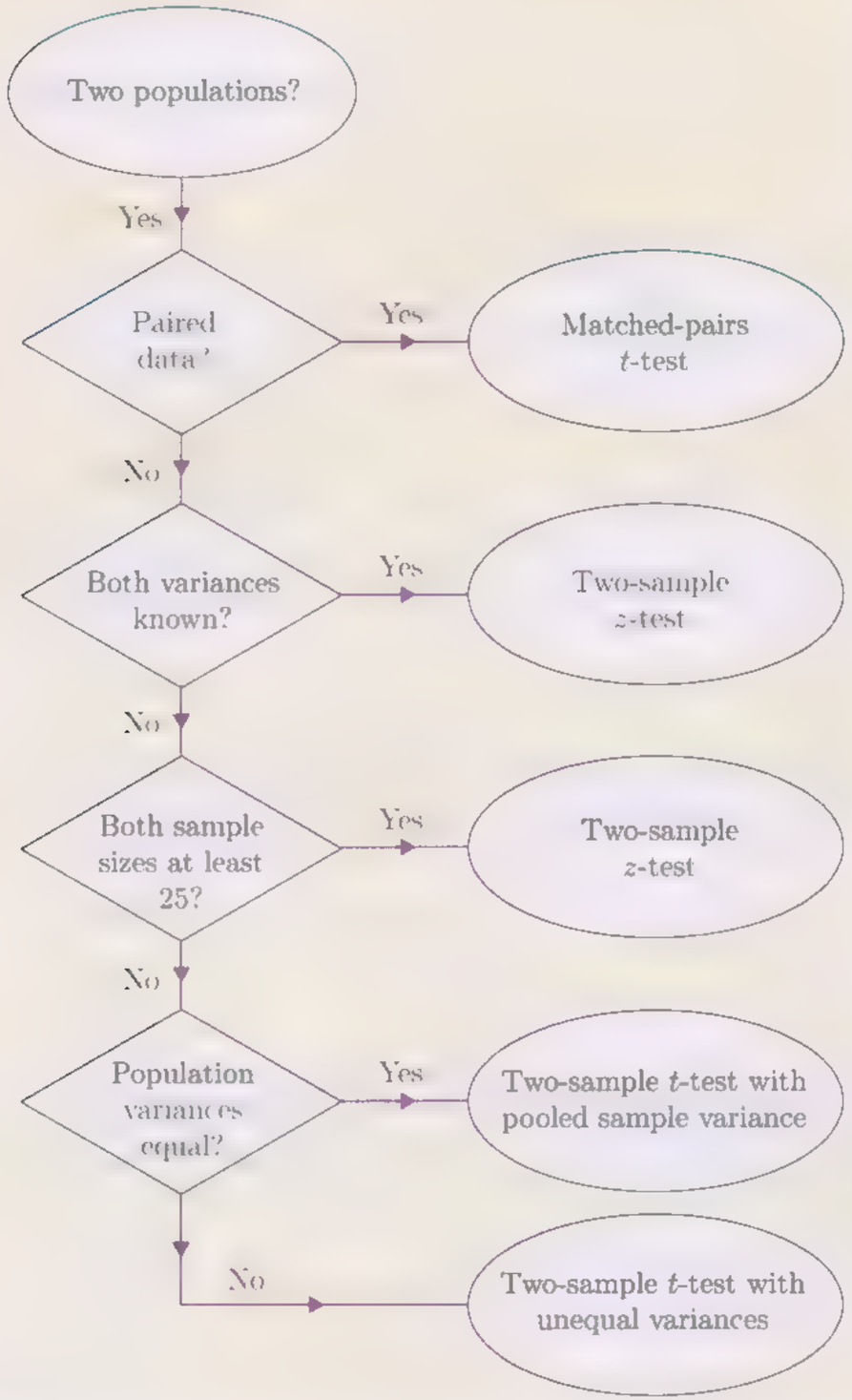
The binomial distribution is the probability distribution of x.

Choosing a test when there is one population



Flow chart for choosing a hypothesis test for inference about a population mean. (It is assumed that observations are random and that the population distribution is approximately normal if the sample size is small.)

Choosing a test when there are two populations



Flow chart for choosing a hypothesis test for inference about the difference between two population means. (Assumptions required for the selected test must also be satisfied.)

Comparing tests

- Null and (two-sided) alternative hypotheses for the z- and t-tests

	H_0	H_1
One-sample tests	$\mu = A$	$\mu \neq A$
Matched-pairs tests	$\mu_d = 0$	$\mu_d \neq 0$
Other two-sample tests	$\mu_A - \mu_B = 0$	$\mu_A - \mu_B \neq 0$

- Estimated standard error (ESE) and test statistic for the z- and t-tests

Test	ESE	Test statistic
1. One-sample z-test	$\dfrac{s}{\sqrt{n}}$	$z = \dfrac{\overline{x} - A}{\text{ESE}}$
2. One-sample t-test	$\dfrac{s}{\sqrt{n}}$	$t = \dfrac{\overline{x} - A}{\text{ESE}}$
3. Two-sample z-test	$\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}$	$z = \dfrac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$
4. Matched-pairs t-test	$\dfrac{s}{\sqrt{n}}$	$t = \dfrac{d}{\text{ESE}}$
5. Two-sample t-test with a common variance	$s_{\text{p}}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}$	$t - \dfrac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$
6. Two-sample t-test with unequal variances	$\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}$	$t = \dfrac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$

The result of the hypothesis test should be stated clearly and conclusions drawn that reflect the setting from which the data came. It is also good practice to state any assumptions that have been made. These will involve the randomness and independence of observations and, for samples of modest size, the assumption that variation in a population is adequately modelled by a normal distribution.

Some useful tables

Interpretation of p-values

(Table 10 in Subsection 5.1 of Unit 6, and reproduced as Table 1 in Subsection 6.2 of the Computer Book)

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

Critical values for the sign test

(Table 8 in Subsection 4.1 of Unit 6)

Sample size	Critical value at the 5% significance level	Sample size	Critical value at the 5% significance level
1	–	21	5
2	–	22	5
3	–	23	6
4	–	24	6
5	–	25	7
6	0	26	7
7	0	27	7
8	0	28	8
9	1	29	8
10	1	30	9
11	1	31	9
12	2	32	9
13	2	33	10
14	2	34	10
15	3	35	11
16	3	36	11
17	4	37	12
18	4	38	12
19	4	39	12
20	5	40	13

Critical values of χ^2

(Extended version of Table 28 in Subsection 4.4 of Unit 8)

Degrees of freedom	Critical values of χ^2 at significance level		Degrees of freedom	Critical values of χ^2 at significance level	
	5%	1%		5%	1%
1	3.841	6.635	21	32.671	38.932
2	5.991	9.210	22	33.924	40.289
3	7.815	11.345	23	35.172	41.638
4	9.488	13.277	24	36.415	42.980
5	11.070	15.086	25	37.652	44.314
6	12.592	16.812	26	38.885	45.642
7	14.067	18.475	27	40.113	46.963
8	15.507	20.090	28	41.337	48.278
9	16.919	21.666	29	42.557	49.588
10	18.307	23.209	30	43.773	50.892
11	19.675	24.725	31	44.985	52.191
12	21.026	26.217	32	46.194	53.486
13	22.362	27.688	33	47.400	54.776
14	23.685	29.141	34	48.602	56.061
15	24.996	30.578	35	49.802	57.342
16	26.296	32.000	36	50.998	58.619
17	27.587	33.409	37	52.192	59.893
18	28.869	34.805	38	53.384	61.162
19	30.144	36.191	39	54.572	62.428
20	31.410	37.566	40	55.758	63.691

5% critical values for a two-sided *t*-test

(Table 2 in Subsection 3.3 of Unit 10)

Degrees of freedom	Critical value (t_c)	Degrees of freedom	Critical value (t_c)
1	12.706	21	2.080
2	4.303	22	2.074
3	3.182	23	2.069
4	2.776	24	2.064
5	2.571	25	2.060
6	2.447	26	2.056
7	2.365	27	2.052
8	2.306	28	2.048
9	2.262	29	2.045
10	2.228	30	2.042
11	2.201	31	2.040
12	2.179	32	2.037
13	2.160	33	2.035
14	2.145	34	2.032
15	2.131	35	2.030
16	2.120	36	2.028
17	2.110	37	2.026
18	2.101	38	2.024
19	2.093	39	2.023
20	2.086	40	2.021

5% critical values for a one-sided *t*-test

(Table 5 in Section 6 of Unit 10)

Degrees of freedom	Critical value (t_c)	Degrees of freedom	Critical value (t_c)
1	6.314	21	1.721
2	2.920	22	1.717
3	2.353	23	1.714
4	2.132	24	1.711
5	2.015	25	1.708
6	1.943	26	1.706
7	1.895	27	1.703
8	1.860	28	1.701
9	1.833	29	1.699
10	1.812	30	1.697
11	1.796	31	1.696
12	1.782	32	1.694
13	1.771	33	1.692
14	1.761	34	1.691
15	1.753	35	1.690
16	1.746	36	1.688
17	1.740	37	1.687
18	1.734	38	1.686
19	1.729	39	1.685
20	1.725	40	1.684

SUP 043738

Cover image: minxlj/www.flickr.com